End User Documentation **CUSTOMER**

Document Version: 1.0 - 2014-07

SAP InfiniteInsight® 7.0

Classification, Regression, Segmentation and Clustering - User Guide

Table of Contents

1 How to Use this Document 4
1.1 Organization of this Document 4
1.2 Which Sections should you Read? 5
1.3 Conventions Used in this Document 6

2 Welcome to this Guide 7
2.1 About this Document 7
2.1.1 Who Should Read this Document 7
2.1.2 Prerequisites for Use of this Document 7
2.1.3 What this Document covers 8
2.2 Before Beginning 8
2.2.1 Files and Documentation Provided with this Guide 8

3 SAP InfiniteInsight® 10
3.1 Introduction 10
3.2 Architecture and Operations 11
3.2.1 User Interfaces 11
3.2.2 Operations 12
3.3 Methodological Prerequisites 15
3.3.1 What is your Business Issue? 15
3.3.2 Is your Data Usable? 15

4 Essential Concepts 16
4.1 Operation of SAP InfiniteInsight®: Overview 17
4.2 Data Sources Supported 18
4.3 Data Set 18
4.3.1 Training Data Set 18
4.3.2 Application Data Set 19
4.4 Cutting Strategies 19
4.4.1 Definition 19
4.4.2 Roles of the Three Sub-Sets 20
4.4.3 Types of Cutting Strategies 20
4.5 Table of Data 24
4.5.1 Definition 24
4.5.2 Synonyms of "Observation" and "Variable" 25
4.5.3 Data Formats 25
4.6 Variables 25
4.6.1 Generic Definition 25
4.6.2 Example 26
4.6.3 Types of Variables 26
4.6.4 Storage Formats 29
4.6.5 Roles of Variables 30
4.7 Models 35
4.7.1 Fundamental Definition 35
4.7.2 Performance of a Model 35
4.7.3 Types of Models 35
4.7.4 Generating a Model 36
4.7.5 Representation of a Model 36
4.7.6 Validating the Model 37
4.7.7 Under what Circumstances is a Model Acceptable? 38
4.7.8 How to Obtain a Better Model 38

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.8 Performance Indicators 39
4.8.1 Indicators Specific to SAP InfiniteInsight® 39
4.8.2 Other Commonly Used Indicators 41
4.8.3 Error Indicators 43
4.9 Profit Type 46
4.9.1 Definition 46
4.9.2 Available Profit Types 46
4.10 Advanced Model Curves 47
4.10.1 ROC 47
4.10.2 Lorenz Curves 48
4.10.3 Density Curves 49
4.10.4 Risk Curves 51

5 InfiniteInsight® Modeler - Regression/Classification 54
5.1 Application Scenario: Enhance Efficiency and Master your Budget using Modeling 54
5.1.1 Presentation 54
5.1.2 Your Objective 54
5.1.3 Your Means 55
5.1.4 Your Approach 56
5.1.5 Your Business Issue 57
5.1.6 Your Solutions 57
5.1.7 Introduction to Sample Files 60
5.1.8 InfiniteInsight® modeling assistant 62
5.2 Creating a Classification Model Using InfiniteInsight® Modeler 66
5.2.1 Step 1 - Defining the Modeling Parameters 66
5.2.2 Step 2 - Generating and Validating the Model 109
5.2.3 Step 3 - Analyzing and Understanding the Model Generated 113
5.2.4 Step 4 - Using the Model 154

6 InfiniteInsight® Modeler - Segmentation/Clustering 197
6.1 Application Scenario: Customize your Communications using Data Modeling 197
6.1.1 Presentation 197
6.1.2 Your Objective 198
6.1.3 Your Approach 198
6.1.4 Your Business Issue 198
6.1.5 Your Solutions 199
6.1.6 Introduction to Sample Files 201
6.1.7 InfiniteInsight® modeling assistant 203
6.2 Creating a Clustering Model Using InfiniteInsight® Modeler 207
6.2.1 Step 1 - Defining the Modeling Parameters 207
6.2.2 Step 2 - Generating and Validating the Model 227
6.2.3 Step 3 - Analyzing and Understanding the Model Generated 232
6.2.4 Step 4 - Using the Model 263

7 Glossary 271

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

1 How to Use this Document

IN THIS CHAPTER

Organization of this Document....4
Which Sections should you Read?....5
Conventions Used in this Document....6

1.1 Organization of this Document

This document is subdivided into six chapters.

This chapter, **Welcome to this Guide**, serves as an introduction to the rest of the guide. This is where you will find information pertaining to the reading of this guide, and information that will allow you to contact us.

Chapter 2, **SAP InfiniteInsight®**, provides an overview of SAP InfiniteInsight®, its architecture and its operation. In addition, it presents two indispensable prerequisite methodologies for using SAP InfiniteInsight® features.

Chapter 3, **Essential Concepts**, introduces the concepts essential to modeling data with the SAP InfiniteInsight®.

The shorter Chapter 4, **General Introduction to Scenarios**, provides a summary of the application scenarios for features InfiniteInsight® Modeler - Regression/Classification and InfiniteInsight® Modeler - Segmentation/Clustering. It also introduces the user interface and the data files used in these scenarios.

Chapters 5 and 6, **Generating Explanatory and Predictive Models with InfiniteInsight® Modeler - Regression/Classification feature and Generating Descriptive Models with InfiniteInsight® Modeler - Segmentation/Clustering feature**, present the InfiniteInsight® Modeler - Regression/Classification feature and the InfiniteInsight® Modeler - Segmentation/Clustering feature of SAP InfiniteInsight®, respectively. These two chapters are organized in the same manner, in two parts:

- The first part introduces a detailed application scenario of that feature,
- The second part introduces the actual use of the feature itself, using the illustration of the corresponding application scenario.

A summary and detailed table of contents located at the beginning of the guide, an index located at the end, and cross-references throughout the document allow you to find the information that you need, quickly and easily.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

1.2 Which Sections should you Read?

Depending on your job profile and your needs, you may choose to read the entire guide, or only certain sections. In either case, it is essential that you read the section concerning the SAP InfiniteInsight® performance indicators (on page 39). These indicators embody one of the most important concepts of SAP InfiniteInsight®: they allow evaluation of the quality and robustness of the models generated.

The following table provides some points of reference to facilitate your use of this guide.

What is your Profile?	Best Use of this Guide
You want to evaluate SAP InfiniteInsight® and your time is tightly budgeted	You could restrict yourself to: 1. Reading the scenario of the feature that interests you (or at least, the summary of that scenario): • **Application Scenario: Enhance Efficiency and Master your Budget using Modeling** • **Application Scenario: Customize your Communications using Data Modeling** 2. Going directly to the relevant section: • **InfiniteInsight® Modeler - Regression/Classification** • **InfiniteInsight® Modeler - Segmentation/Clustering**
▪ You want to be guided step-by-step through SAP InfiniteInsight® ▪ You have had only limited hands-on experience in data modeling	Read all sections of the guide through at least once, in the order in which they are presented. In both cases, ensure that you have a complete grasp of the essential concepts relating to the use of SAP InfiniteInsight®, by reading chapter **Essential Concepts** (on page 16). These concepts are essential both for the use of SAP InfiniteInsight® features and for analysis of the results obtained.
You have had **significant experience** in data modeling	You could limit yourself to: 1. Verifying that you are familiar with the terminology used by InfiniteInsight, by examining the contents of chapter **Essential Concepts** (on page 16), in the detailed table of contents. 2. Reading the summary of the scenario of the feature that interests you. • **Application Scenario: Enhance Efficiency and Master your Budget using Modeling** • **Application Scenario: Customize your Communications using Data Modeling** 3. Going directly to the relevant section: • **InfiniteInsight® Modeler - Regression/Classification** • **InfiniteInsight® Modeler - Segmentation/Clustering**
▪ You have previously taken a SAP InfiniteInsight® training seminar ▪ You are already a SAP InfiniteInsight® user	You can: ▪ Follow the application scenarios for a "review " of the features that interest you. • **Application Scenario: Enhance Efficiency and Master your Budget using Modeling** • **Application Scenario: Customize your Communications using Data Modeling** ▪ Use this document as a reference text, consulting it as required. In this case, the detailed table of contents and the index will be valuable tools, helping you find the information that you seek.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

1.3 Conventions Used in this Document

To facilitate reading, certain publishing conventions are applied throughout this guide. These are presented in the following table.

The following information items...	Are presented using...	For example...
Graphical interface features and file names	*Arial bold*	Click *Next*
The titles of particularly useful sections	Garamond italicized bold	See Operations
The titles of procedures		**To Select the Target Variable**
The titles of sections specific to the scenario presented in this guide		**For this Scenario**

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2 Welcome to this Guide

IN THIS CHAPTER

About this Document 7
Before Beginning 8

2.1 About this Document

2.1.1 Who Should Read this Document

This document is addressed to people who want to **evaluate** or **use** SAP InfiniteInsight®.

2.1.2 Prerequisites for Use of this Document

Use of this guide does not require any prior expertise in statistics or databases.

SAP InfiniteInsight® features are developed using cutting-edge technologies and while they call on complex, innovative statistical techniques, they are still straightforward and quick to use: they put powerful Data Mining techniques within the reach of non-expert users.

For more technical details regarding SAP InfiniteInsight®, please contact us. We will be happy to provide you with more technical information and documentation.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2.1.3 What this Document covers

This document introduces you to the basic concepts underpinning SAP InfiniteInsight®, and the main functionalities of the InfiniteInsight® Modeler - Regression/Classification and InfiniteInsight® Modeler - Segmentation/Clustering features. Using two application scenarios, you can create your first models with confidence.

This document is the primary guide to the two SAP InfiniteInsight® features described in the following table.

The feature...	Allows you to...	Example...
InfiniteInsight® Modeler - Regression/Classification	Understand and predict a phenomenon	You work for an automobile manufacturer and wish to send a promotional mailing to your prospects. InfiniteInsight® Modeler - Regression/Classification allows you to: ▪ Understand why previous prospects responded to such a mailing, ▪ Predict the response rate to such a mailing sent to new prospects.
InfiniteInsight® Modeler - Segmentation/Clustering	Describe a data set, by breaking it down into homogeneous data groups, or clusters	Your firm is in the process of bringing products A and B to market. InfiniteInsight® Modeler - Segmentation/Clustering allows you to: ▪ Regroup your customers into several homogeneous groups, ▪ Understand the behavior of each of these groups with respect to products A and B.

2.2 Before Beginning

2.2.1 Files and Documentation Provided with this Guide

Sample Data Files

The evaluation version and the registered version of SAP InfiniteInsight® are supplied with sample data files. These files allow you to take your first steps using various features of SAP InfiniteInsight®, and evaluate them.

During installation of SAP InfiniteInsight®, the sample files are registered in the folder: `C:\Program Files\SAP InfiniteInsight\InfiniteInsightVx.y.z\\Samples\Census\`.

The following table describes those files.

File Name	Description	When is it Used?
Census01.csv	Data file	This file is used for both application scenarios used in this manual.
desc_census01.csv	Description file for the *Census01.csv* file	This file is used for both application scenarios used in this manual.

To obtain a detailed description of the `Census01.csv` file, see Introduction to Sample Files (see page 60).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Documentation

Full Documentation

Complete documentation is included with SAP InfiniteInsight®. This documentation covers:

- The operational use of SAP InfiniteInsight® features,
- The architecture and integration of the SAP InfiniteInsight® API,
- The SAP InfiniteInsight® Java graphical user interface.

Contextual Help

Each screen in the InfiniteInsight® modeling assistant is accompanied by contextual help that describes the options presented to you, and the concepts required for their application.

☑ To Display the Contextual Help

In the *Help* menu located at the upper left corner of InfiniteInsight® modeling assistant, click the option *Help*.
You can also directly press `F1` on your keyboard to display quickly the contextual help.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

3 SAP InfiniteInsight®

IN THIS CHAPTER

Introduction 10
Architecture and Operations 11
Methodological Prerequisites 15

3.1 Introduction

SAP InfiniteInsight® platform has be developed in order to provide the ideal Data Mining solution for modeling your data as easily and rapidly as possible, while maintaining relevant and readily interpretable results. Thanks to SAP InfiniteInsight®, you will transform your data into knowledge in order to make timely strategic and operational decisions.

SAP InfiniteInsight® places the latest Data Mining techniques within reach of any non-expert user. SAP InfiniteInsight® allows you to access many data source formats and to generate explanatory and predictive models, as well as descriptive models in a semi-automated manner, extremely rapidly.

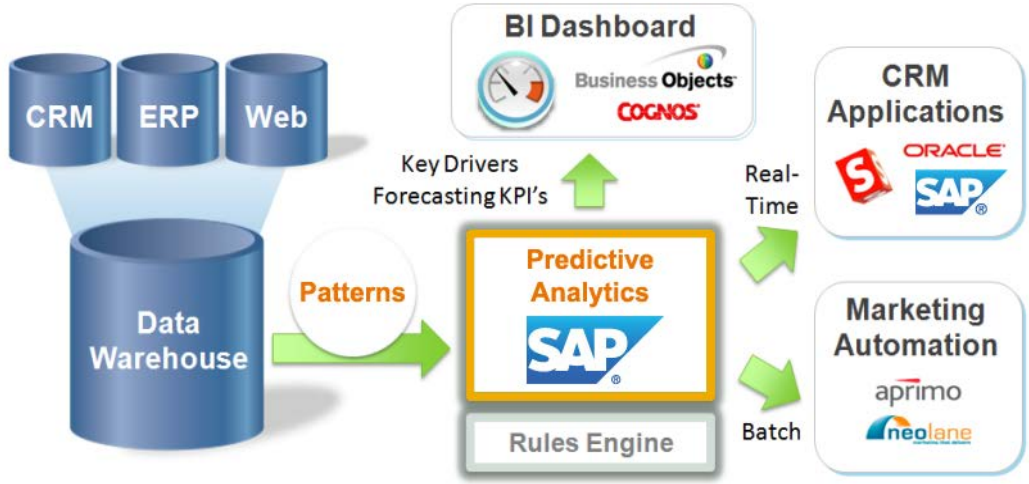
With SAP InfiniteInsight®, you can concentrate on high value-added activities such as analysis of the results of data modeling, and decision-making.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

3.2 Architecture and Operations

The figure below illustrates the general architecture of SAP InfiniteInsight®.

This section provides an introduction to the elements of this architecture like the various types of **interfaces** that allow you to use SAP InfiniteInsight®.



3.2.1 User Interfaces

Three Types of User Interface

Three types of interfaces allow you to use the features of SAP InfiniteInsight®:

- **Graphical user interface**,
- **Command interpreter**,
- **API (***Application Programming Interface***) controls**.

Graphical User Interfaces

The **graphical user interface** is aimed primarily at "end users" or "non-expert users". It provides access to the **InfiniteInsight® modeling assistant**, which allow you to use SAP InfiniteInsight® features and model your data very easily. In addition, it provides plotted output to facilitate viewing and interpretation of the results of modeling.

The graphical user interface provided with SAP InfiniteInsight® is the **InfiniteInsight® modeling assistant** interface, developed in Java on the CORBA API (*Application Programming Interface*), which operates on any platform (Windows, UNIX, among others). This interface is provided as an example. In addition, with the *Application Programming Interface* furnished with SAP InfiniteInsight®, you can develop your own graphical interfaces.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

The KxShell Command Interpreter

The **KxShell** command interpreter allows you to use SAP InfiniteInsight® by typing commands or executing scripts containing several commands.

The command interpreter is an example of development based on the C++ API. Like any other API, it may be used to integrate SAP InfiniteInsight® with other applications or program packages.

Control API

The **Control API** (*Application Programming Interface*) is aimed primarily at developers, or users with programming experience. This *Application Programming Interface* is used to access the complete range of functionalities and the most fine-grained parameterization of SAP InfiniteInsight® features. In addition, it allows customized integration of SAP InfiniteInsight® features with other applications or program packages.

Three APIs are provided with SAP InfiniteInsight®:

- A **COM/DCOM** API, usable over Microsoft platforms,
- A **CORBA** API, usable over all client/server platforms,
- A **C++** API, usable over all standalone platforms.

3.2.2 Operations

The operation of SAP InfiniteInsight® may be subdivided into four phases.

- Phase 1 - Data access (on page 12)
- Phase 2 - Data manipulation and preparation (on page 13)
- Phase 3 - Data modeling (on page 14)
- Phase 4 - Model presentation and deployment (on page 14)

Phase 1: Data Access

SAP InfiniteInsight® accepts many types of data sources:

- "Flat files" such as .csv files, files of text tables and other files of type text.
- ODBC-compatible sources such as Oracle, SQL Server or IBM DB2 databases.

In addition, the **C Data Access** *Application Programming Interface* allows you to connect proprietary format sources, such as industrial sensor streams.

In most cases, and particularly if you are using SAP InfiniteInsight® features via a graphical interface, you never have to concern yourself with the data access process. Data access is accomplished in a semi-transparent manner: from the graphical user interface, you need only select the data source format to be used ("flat files" or ODBC-compatible data sources) and specify the location of the data file. The C Data Access *Application Programming Interface* is helpful to developers who want to write access code for proprietary format databases.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

The InfiniteInsight® Access Feature

The **InfiniteInsight® Access** (formerly known as KAA) feature allows reading SAS data and writing the scores obtained with an SAP InfiniteInsight® model into a SAS table.

The following formats are currently supported:

- SAS files version 6 under windows & unix
- SAS 7/8 under windows & unix
- SAS Transport Files

You can access directly a SAS table with the SAP InfiniteInsight® interface simply by selecting the format of the file to analyze. Once you have built your model with SAP InfiniteInsight®, you can generate a SAS table containing the model application results (scores, probability, cluster number, predicted value, ...). The SAP InfiniteInsight® interface allows you to select the output format. The generated SAS table is automatically integrated in SAS information system.

Phase 2: Data Manipulation and Preparation

The InfiniteInsight® Explorer - Sequence Coding and InfiniteInsight® Explorer - Event Logging features are data manipulation and preparation features. They are used to encode data in a robust and semi-automatic manner, making them available for use by all analytical features of the SAP InfiniteInsight®. The use of these features is transparent for the final user: all data processing is performed in a completely automatic manner.

InfiniteInsight® Explorer - Event Logging (formerly known as KEL) aggregates events into periods of time. It allows integrating transactional data with demographic customer data. It is used in cases when the raw data contains static information such as age, gender or profession of an individual, and dynamic variables, such as spending patterns or credit card transactions. Data is automatically aggregated within user defined periods without programming SQL or changing database schema. InfiniteInsight® Explorer - Event Logging combines and compresses this data to make it available to other SAP InfiniteInsight® features.

InfiniteInsight® Explorer - Sequence Coding (formerly known as KSC) aggregates events into a series of transitions. For example a customer click-stream from a Web site can be transformed into a series of data for each session. Each column represents a specific transition from one page to another. Similar to InfiniteInsight® Explorer - Event Logging these new columns of data can be added to existing customer data and are made available to other SAP InfiniteInsight® features for further processing.

InfiniteInsight® Modeler - Data Encoding (formerly known as K2C) automatically prepares and transforms data into a format suitable for use in the SAP InfiniteInsight®. InfiniteInsight® Modeler - Data Encoding translates nominal and ordinal variables, automatically fills in missing values and detects out of range data. In addition, this feature contributes significantly to the robustness of the models generated by the SAP InfiniteInsight® engine, by providing a robust data encoding.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Phase 3: Data Modeling

Thanks to the statistical techniques and information technologies upon which the InfiniteInsight® Modeler - Regression/Classification, InfiniteInsight® Modeler - Segmentation/Clustering and InfiniteInsight® Modeler – Time Series features were built, these features require only an extremely short modeling time to generate relevant and robust analytical models of your data.

InfiniteInsight® Modeler - Regression/Classification (formerly known as K2R) generates explanatory and predictive models. The models generated by Classification/Regression explain and predict a phenomenon, or business question, by a function of the analyzed data set, the explanatory variables. The models generated are calculated using a regression and classification algorithm. This polynomial regression is a proprietary algorithm using Vapnik's SRM (*Structural Risk Minimization*) principle to calculate the parameters.

InfiniteInsight® Modeler - Segmentation/Clustering (formerly known as K2S) generates descriptive models, which means a function to regroup cases in a data set into a number of clusters with similar behavior toward a business question.

InfiniteInsight® Modeler – Time Series (formerly known as KTS) lets you build predictive models from data representing time series. Thanks to InfiniteInsight® Modeler – Time Series models, you can:

- Identify and understand the phenomenon represented by your time series.
- Forecast the evolution of time series in the short and medium term, that is, predict their future values.

Phase 4: Model Presentation and Deployment

Once the models have been generated, model performance indicators, plots and modeling reports in HTML format facilitate viewing and interpretation of the data modeling results.

Once the models have been validated, you can apply them to:

- One or more specific observations taken from your database (*Simulation mode*),
- A new, complete data set or application data set (*Batch mode*).

To facilitate deployment and integration of the models, the code corresponding to each model can also be generated in the programming language. The InfiniteInsight® Scorer feature, which is responsible for generation of this code, is described below.

The InfiniteInsight® Scorer Feature

The **InfiniteInsight® Scorer** feature (formerly known as KMX) generates code in the following languages: C, XML, AWK, HTML, SQL, PMML2, SAS, or JAVA corresponding to a model generated by SAP InfiniteInsight®.

In this form, the model may be integrated into any application that supports the aforementioned languages.

The generated codes allow the SAP InfiniteInsight® models to be integrated within any given application or software package, or to be applied directly to the data without requiring SAP InfiniteInsight® environment.

Caution

Code generation is only available for models using the following features: **InfiniteInsight® Modeler - Data Encoding**, **InfiniteInsight® Modeler - Regression/Classification**, **InfiniteInsight® Modeler - Segmentation/Clustering**.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

3.3 Methodological Prerequisites

Before modeling your data using the SAP InfiniteInsight®, you should :

- State a business issue that you want to solve,
- Possess a data set representing this issue in the form of a set of observations.

3.3.1 What is your Business Issue?

All of the SAP InfiniteInsight® features are a response to the same requirement: they allow **supervised data analysis**. The term "supervised" means that the data analysis does not occur completely independently, but always as a function of a particular issue: your business issue!

Consider the database that contains information about your customers. An analysis that grouped your customers into homogeneous groups independently of your input would be of little interest. On the other hand, an analysis that grouped them as a function of a variable such as "mean business revenues earned from this customer each year" would offer significant interest. You would learn the characteristic profiles of the customers that bring you the most money. Then, you can develop strategies to better influence your customers according to their characteristic profiles.

To recap, the prerequisite step before using SAP InfiniteInsight® consists of **identifying and formulating your business issue**.

3.3.2 Is your Data Usable?

Once your business issue has been identified and formulated, you need to have data on hand that will permit an answer to be found. We will not expound at length about the information value associated with data. This depends on your data collection and extraction processes and tools, and not SAP InfiniteInsight® features. On the other hand, in order for your data to be usable by SAP InfiniteInsight®, the following five conditions must be met:

- You must have **a sufficiently large volume of data** to be able to build a valid model, that is, in order for the model to be both relevant and robust. An analytical model that is generated from a data set of 50 lines may have low generalization capacity, and contain low informative value. We can advise you on the issues of data volume.
- Your data set must contain a **target variable**, that will allow expression of your business issue within SAP InfiniteInsight®.
- **The target variable must be known for each observation of the training data set**. To express this another way, no target variable values may be missing over the range of the entire training data set.
- The **data source format** must be supported by SAP InfiniteInsight®.
- Your data must be presented in the form of a **single table of data**, except in instances where you are using the InfiniteInsight® Explorer - Event Logging or InfiniteInsight® Explorer - Sequence Coding features.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4 Essential Concepts

This section introduces the essential concepts relating to use of SAP InfiniteInsight®.

All concepts are introduced and appear in boldface in the section Operation of SAP InfiniteInsight® **Overview**, which provides a general description of the use of SAP InfiniteInsight® in the model generation process.

IN THIS CHAPTER

Operation of SAP InfiniteInsight®: Overview 17
Data Sources Supported 18
Data Set 18
Cutting Strategies 19
Table of Data 24
Variables 25
Models 35
Performance Indicators 39
Profit Type 46
Advanced Model Curves 47

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.1 Operation of SAP InfiniteInsight®: Overview

SAP InfiniteInsight® allows you to perform supervised Data Mining, that is, to transform your data into knowledge, then into action, as a function of a domain-specific business issue.

SAP InfiniteInsight® supports various **formats of source data** (**flat files**, **ODBC-compatible sources**). In order to be usable by SAP InfiniteInsight® features, the data set to be analyzed must be presented in the form of a single **table of data** (on page 24), except in instances where you are using the Event Logging or Sequence Coding features of InfiniteInsight® Explorer.

To use SAP InfiniteInsight® features, you must have a **training data set** available that contains the target variable with all its values defined. Then, you can apply the model generated using the training data set to **one or more application data sets**.

The training data set is cut into **three data sub-sets** for **estimation**, **validation** and **testing**, using a **cutting strategy** (on page 19).

The different **types of variables** (on page 26): **continuous**, **ordinal** and **nominal** are next encoded by the Data Encoding feature of InfiniteInsight® Modeler, or by the Event Logging and Sequence Coding features in the case of dynamic data. Before generating the model, you must:

- Describe the data. A utility integrated with SAP InfiniteInsight® allows you to generate a description of the data set to be analyzed, automatically. You need only validate that description, verifying that the **type** and **storage format** of each variable were identified correctly.
- Define the **role of variables** contained in the data set to be analyzed. You may select one or more variables as **target variables**. These are the variables that corresponds with your business issue. The other variables of the table of data are considered to be **explanatory variables**: they allow calculation of the value of the target variable in a given context. They may also be used as **weight variables**.

For more information about the role of each SAP InfiniteInsight® features, see Operations (see page 12).

You can then generate **models** (on page 35), capable of either explaining and predicting a phenomenon, or describing a data set, in both cases, as a function of the previously defined target variable. This phase is called the training phase.

Once the models have been generated, you can view and interpret their relevance and robustness using:

- **Performance indicators** (on page 39): the predictive power, which is the quality indicator, and the prediction confidence , which is the robustness indicator,
- A variety of plots, including the **profit curve** plot.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.2 Data Sources Supported

In the standard version, SAP InfiniteInsight® supports the following data sources:

- **Flat files** (text files) in which the data are separated by a delimiter, such as commas in .csv (Comma Separated Value) format file. For instance, the sample file *Census01.csv*, used for the InfiniteInsight® Modeler - Regression/Classification and InfiniteInsight® Modeler - Segmentation/Clustering application scenarios, is a .csv file.
- **ODBC-compatible data sources**.
- if your license allows it you can also use **SAS files**.

An API also allows you to interface SAP InfiniteInsight® with any other application (SPSS, Microsoft Excel, for example), and gain access to any other data source. You must develop a specific .dll script for each new source.

> **Note**
>
> For information about data formatting, and specifically, for the list of supported ODBC-compatible sources, see the document **Data Modeling Specification**.

4.3 Data Set

To use SAP InfiniteInsight® features, you must have a **training data set** available that contains the target variable with all its values defined. Then, you can apply the model generated using the training data set to **one or more application data sets**.

4.3.1 Training Data Set

A **training data set** is a data set used for generating a model. In this set, the values of the **target variable** (on page 31) – or variable corresponding to your business issue – are known. By analyzing the training data set, SAP InfiniteInsight® features will generate a model that allows explanation of the target variable, based on the explanatory variables.

To allow validation of the model generated, the training data set is cut into three sub-sets using a **cutting strategy** (on page 19).

The training data set may correspond to either a complete population section of your database or a sample extracted from this population. The choice depends on the type of study to be performed, the tools used and the budget allocated to the study.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.3.2 Application Data Set

An **application data set** is a data set to which you apply a model. This data set contains an unknown target variable for which you want to know the value.

The model applied to the application data set must have been previously generated from a training data set. The application data set must contain exactly the same information structure as the corresponding training data set, that is:

- The same number of variables,
- The same types of variables,
- The same order of presentation of these variables.

Caution

The application data set must contain a target variable that corresponds to that of the training data set. This point is true for all instances, even if the values of this target variable are empty. When these values are defined, they may serve to detect any possible deviant observations (outliers).

4.4 Cutting Strategies

4.4.1Definition

A **cutting strategy** is a technique that allows decomposition of a training data set into three distinct sub-sets:

- An **estimation sub-set**,
- A **validation sub-set**,
- A **test sub-set**.

This cutting allows for **cross-validation** of the models generated.

There are nine types of cutting strategies available within SAP InfiniteInsight®.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.4.2 Roles of the Three Sub-Sets

The following table defines the roles of the three data sub-sets obtained using cutting strategies.

The data-set	Is used to...
Estimation	Generate different models. The models generated at this stage are hypothetical.
Validation	Select the best model among those generated using the estimation sub-set, which represents the best compromise between perfect quality and perfect robustness.
Test	Verify the performance of the selected model on a new data set.

To understand the role of cutting strategies in the model generation process, see the figure **Generating a Model** (see page 36).

4.4.3 Types of Cutting Strategies

To generate your models, there are two types of cutting strategies that you may use:

- The **customized cutting strategy**,
- The **automatic cutting strategies**.

Customized Cutting Strategy

Definition

The **customized cutting strategy** allows you to define your own data sub-sets. To use this strategy, you must have prepared (before opening SAP InfiniteInsight® features) three sub-sets: the estimation, validation and test sub-sets.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

How to Use this

Before opening the SAP InfiniteInsight®, cut your initial data file into three files of the size of your choice. For example:

- The first file may contain the first 1,500 observations, or lines, of your initial data file,
- The second file, observations 1,501 to 3,000,
- The third file, observations 3,001 to 5,000.

Caution

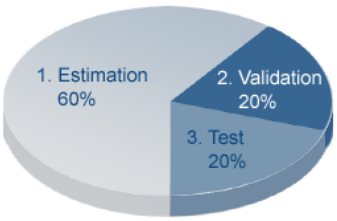
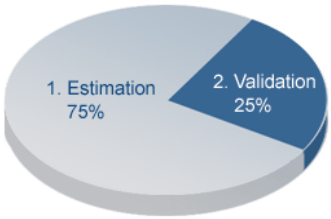
The *customized* cutting strategy is risky in the instance of an initial data file in which the data have been sorted. In this case, the first lines will not be representative of the overall set of data contained in the first file. To avoid this type of bias, do not forget to mix up your data prior to analysis.

Seven Automatic Cutting Strategies

Background

With the exception of the customized cutting strategy, cutting strategies are automatic. **Automatic cutting strategies** operate upon a single data file, which constitutes your initial data set.

Automatic cutting strategies always cut the initial data set into the same proportions. The following table details the proportions attributed to each data set depending on the presence of a test data set.

Automatic Cutting Strategies with Test	Automatic Cutting Strategies without Test
▪ 3/5 of the data are used in the estimation sub-set, ▪ 1/5 of the data are used in the validation sub-set, ▪ 1/5 of the data are used in the test sub-set. 	▪ 3/4 of the data are used in the estimation sub-set, ▪ 1/4 of the data are used in the validation sub-set. 

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Random

The **Random** cutting strategy distributes the data of the initial data set in a random manner between the three sub-sets, estimation, validation and test.

Random with Test at the End

The **Random with test at the end** cutting strategy distributes:

- 4/5 of the initial data set in a random manner in the two sub-sets estimation and validation, 3/5 being distributed in the estimation data sub-set and 1/5 in the validation data sub-set.
- The final 1/5 of the initial data set is sent directly into the test sub-set.

This is a useful strategy in cases where:

- Your database corresponds to a well-defined evolution because of the way it was built, which may mean, for example, that the data is in chronological order,
- You may wish to take this order into account when generating your model.

For example, imagine that:

- New customers are added every month to your database,
- You know that the data sets to which you apply the model will, once generated, have a better chance of resembling the most recent section of your database, that is, the section that contains the most recent customers entered.

Using the *Random with test at the end* cutting strategy, you decide to test the model generated on that section of your database that is most likely to resemble the state of your future application data sets.

Random Without Test (Default Strategy)

The **Random without test** strategy is the cutting strategy suggested as the default setting. It distributes the whole initial data set in a random manner to the two sub-sets of estimation and validation:

- 3/4 of the initial data set are distributed to the estimation sub-set,
- 1/4 to the initial data set are distributed to the validation sub-set.

As no test sub-set is used, all the data from your training data set can be used for sub-sets of estimation and validation. This can lead to a model with a better quality and robustness.

Periodic

The **Periodic** cutting strategy is implemented by following this distribution cycle:

1. Three lines of the initial data set are distributed to the estimation sub-set.
2. One line is distributed to the validation sub-set.
3. One line is distributed to the test sub-set.
4. Distribution begins again at step *1*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Periodic with Test at the End

The **Periodic with test at the end** strategy distributes:

- 4/5 of the initial data set in a periodic manner to the two sub-sets of estimation and validation, 3/5 being distributed in the estimation data sub-set and 1/5 in the validation data sub-set 3/5 being distributed.
- The final 1/5 of the initial data set is sent as a block of data to the test sub-set.

In other words, this strategy follows this distribution cycle:

1. Three lines of the first 4/5 of the initial data set are distributed to the estimation sub-set.
2. One line of the first 4/5 of the initial data set is distributed to the validation sub-set.
3. a. If the entire 4/5 of the initial data set is not yet distributed, distribution operations begin again at step *1*.
   b. If the entire 4/5 of the initial data set has been distributed, distribution operations go to step *4*.
4. The final 1/5 of the initial data set is sent as a block of data to the test sub-set.

Periodic Without Test

The **Periodic without test** strategy distributes the whole initial data set in a periodic manner to the two sub-sets of estimation and validation:

- 3/4 of the initial data set are distributed to the estimation sub-set,
- 1/4 to the initial data set are distributed to the validation sub-set.

In other words, this cutting strategy is implemented by following this distribution cycle:

1. Three lines of the initial data set are distributed to the estimation sub-set.
2. One line is distributed to the validation sub-set.
3. Distribution begins again at step *1*.

As no test sub-set is used, all the data from your training data set can be used for sub-sets of estimation and validation. This can lead to a model with a better quality and robustness.

Sequential

The **Sequential** strategy cuts the initial data set into three blocks, corresponding to the usual cutting proportions:

- The lines corresponding to the first 3/5 of the initial data set are distributed as a block to the estimation data set,
- The lines corresponding to the next 1/5 of the initial data set are distributed as a block to the validation data set,
- The lines corresponding to the final 1/5 of the initial data set are distributed as a block to the test data set.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Sequential Without Test

The **Sequential without test** strategy cuts the initial data set into two blocks:

- The lines corresponding to the first 3/4 of the initial data set are distributed as a block to the estimation data set,
- The lines corresponding to the next 1/4 of the initial data set are distributed as a block to the validation data set.

As no test sub-set is used, all the data from your training data set can be used for sub-sets of estimation and validation. This can lead to a model with a better quality and robustness.

4.5 Table of Data

4.5.1 Definition

A **table of data** is a data set presented in the form of a two-dimensional table.

In this table:

- Each row represents an observation to be processed, such as "American individual" in the sample file *Census01.csv*.
- Each column represents a variable that describes observations, such as the "age" or the "gender" of individual Americans.
- Each cell, the intersection of a column and a row, represents the value of the variable in the column, for the observation in that row.

The following table is an example of a table of data.

Observations	Variable 1	Variable 2	Variable 3
Observation a	Value a1	Value a2	Value a3
Observation b	Value b1	Value b2	Value b3
...	...	...	...
Observation n	Value n1	Value n2	Value n3

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.5.2 Synonyms of "Observation" and "Variable"

Depending upon your profile and your area of expertise, you may be more familiar with other terms that refer to observations (in rows) and variables (in columns) when using tables of data.

The following table presents such terms, or synonyms.

Terms equivalent to the term "Observation"	Terms equivalent to the term "Variable"
Row	Column
Record	Attribute
Table	Field
Event	Property
Instance	-
Example	-

4.5.3 Data Formats

Whatever the data source used, the following two constraints must be accommodated:

- The data must be represented in the form of a **single table**, except in instances where you are using the InfiniteInsight® Explorer - Event Logging or InfiniteInsight® Explorer - Sequence Coding features.
- The target variable must be defined for each observation in the table. In the sample file *Census01.csv*, the variable "class" has been be defined for each individual.

> **Note**
>
> For information about data formatting, and specifically, for the list of supported ODBC-compatible sources, see the document **Data Modeling Specification**.

4.6 Variables

4.6.1Generic Definition

A **variable** corresponds to an attribute which describes the observations stored in your database. In SAP InfiniteInsight® features, a variable is defined by:

- Type,
- Storage format,
- Role.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.6.2 Example

In a database containing information about your customers, the "name" and "address" of those customers are examples of variables.

4.6.3 Types of Variables

There are three types of variables:

- Continuous variables,
- Ordinal variables,
- Nominal variables.

Continuous Variables

Definition

Continuous variables are variables whose values are **numerical**, **continuous** and **sortable**. Arithmetic operations may be performed on these values, such as determination of their sum or their mean.

Example

The variable "salary" is a numerical variable, but in addition, is also a **continuous variable**. It may, for instance, take on the following values: "$1,050", "$1,700", or "$1,750". The mean of these values may be calculated.

Continuous Variables and Modeling

During modeling, a continuous variable may be grouped into significant discrete bins.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Ordinal Variables

Definition

Ordinal variables are variables with **discrete** values, that is, they belong to categories, and they are **sortable**. Ordinal variables may be:

- **Numerical**, meaning that its values are numbers. They are therefore ordered according to the natural number system (0, 1, 2, and so on).
- **Textual**, meaning that its values are character strings. They are therefore ordered according to alphabetic conventions.

Example

The variable "school grade" is an **ordinal variable**. Its values actually belong to definite categories and can be sorted. This variable can be :

- numerical, if its values range between "0" and "20",
- textual , if its values are A, B, C, D, E et F.

Caution

A variable "assessment" which values are "good", "average", and "bad" cannot be directly treated as an ordinal variable by SAP InfiniteInsight® features. The values would be sorted in alphabetical order ("average","bad", "good") and not according to their meaning. When a nominal variable order is important, the variable must be encoded, in letters or numbers, before it can be used by SAP InfiniteInsight®.

Nominal Variables

Definition

Nominal variables are variables whose values are **discrete**, that is, belong to categories, and are **not sortable**.

Nominal variables may be:

- **Numerical**, meaning that its values are numbers.
- **Textual**, meaning that its values are character strings.

Caution

Binary variables are considered nominal variables.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Example

The variable "zip code" is a **nominal variable**. The set of values that this variable may assume ("10111", "20500", "90210", for example) are clearly distinct, non-ranked categories, although they happen to be represented by numbers.

The variable "eye color" is a **nominal variable**. The set of values that this variable may assume ("blue", "brown", "black", for example) are clearly distinct, non-ordered categories, and are represented by character strings.

Nominal Variables and Modeling

During modeling, the values of the categorical variables are regrouped into homogeneous categories. These categories are then ordered as a function of their relative contribution with respect to the values of the target variable.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.6.4 Storage Formats

To describe the data, SAP InfiniteInsight® uses four types of storage formats:

- *date*,
- *datetime*,
- *number*,
- *integer*,
- *string*.

The following table describes these storage formats.

The storage format...	Is used to describe variables when their values correspond to...	For instance...
date	Dates expressed in the following formats: ▪ YYYY-MM-DD ▪ YYYY/MM/DD	▪ "2001-11-30" ▪ "1999/04/28"
datetime	Dates and times expressed in the following formats: ▪ YYYY-MM-DD HH:MM:SS ▪ YYYY/MM/DD HH:MM:SS	▪ "2001-11-30 14:08:17" ▪ "1999/04/28 07:21:58"
number	Figures, or numerical values on which operations may be performed	▪ The variable "salary", in US dollars: "1000.00", "1593" and "2000.54" ▪
integer	Figures, or numerical integer values on which operations may be performed	▪ The variable "age", in years: "21", "34" and "99"
string	Alphanumeric character strings	▪ The variable "family name": "Lake", "Martin" and "Miller" ▪ The variable "occupation": "professor", "engineer" and "translator" ▪ The variable "telephone": "800 555 1234" and "800 555 4321"

Note

A variable that has numbers for values is not forced to be described using the *number* storage format. For instance, the variables "telephone" and "zip code" may instead be described using the *string* storage format, because no arithmetic operations that make any sense can be performed on these values. Similarly, a variable that will be used as an observation identification code in a table, and does not comply with supported number formats may be described using the *string* storage format.

Caution

For *number* storage formats, the decimal separator used must be a decimal point, and not a comma. So, the value "6.5" may be processed, while "6,5" will not be processed.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Date and Datetime Variables: Automatically Generated Variables

When your data set contains date or datetime variables, the feature Date Coder automatically extracts date information. The Date Coder is able to extract the following temporal information.

For date or datetime variables:

Temporal Information	Represents...	Generated Variable Name
Day of week	the day of week according to the ISO disposition (Monday=0 and Sunday=6)	`<OriginalVariableName>_DoW`
Day of month	the day of month (1 to 31)	`<OriginalVariableName>_DoM`
Day of year	the day of the current year (1 to 366)	`<OriginalVariableName>_DoY`
Month of quarter	the month of the quarter (January, April, July and October = 1, February, May, August and November = 2, March, June, September and December = 3)	`<OriginalVariableName>_MoQ`
Month of year	the month (1 to 12)	`<OriginalVariableName>_M`
Year	the year	`<OriginalVariableName>_Y`
Quarter	the quarter of the year (January to March = 1, April to June = 2, July to September = 3, October to December = 4)	`<OriginalVariableName>_Q`

From datetime variables:

Temporal Information	Represents...	Generated Variable Name
Hour	the hour	`<OriginalVariableName>_H`
Minute	the minute	`<OriginalVariableName>_Mi`
Second	the second	`<OriginalVariableName>_S`
µ seconds	the micro-second	`<OriginalVariableName>_mu`

The generated variables will appear in the model debriefing panels listing variables, such as the *Contributions by Variable*, the *Category Significance*, the *Statistical Reports*, as well as in the automatic variable selection feature.

4.6.5 Roles of Variables

In data modeling, variables may have three roles. They may be:

- Target variables,
- Explanatory variables,
- Weight variables.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Target Variable

Definition

A **target variable** is the variable that you seek to explain, or for which you want to predict the values in an application data set. It corresponds to your domain-specific business issue.

When the target variable is a binary variable, the SAP InfiniteInsight® considers that the target value, or **target category**, of this variable (that is, the value that is the object of the analysis) to be the least frequently occurring value in the training data set. Imagine that a training data set containing the customer information of a company contains the target variable "responded to my mailing". This target variable may take the values "Yes" or "No". If the value "Yes" is the least frequent value (for instance, if 40% of referenced customers responded to the mailing), the SAP InfiniteInsight® considers that value to be the target category of the target variable.

Synonyms

Depending upon your profile and your area of expertise, you may be more familiar with one of the following terms to refer to target variables:

- Variables to be explained,
- Dependent variables,
- Output variables.

These terms are synonyms.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Example

Your company is marketing two products A and B.

You have a database which contains references to:

- 1,500 of your customers. You know which product, A or B, each customer has purchased.
- 10,000 prospects. You want to know which product each customer is likely to purchase.

The variable "product purchased" is your target variable: it corresponds to your business issue. It is:

- Known for all values of the training data set (in our example, the customers),
- Not known for the values of the application data set (in our example, the prospects).

SAP InfiniteInsight® features allow you to model that target variable, and thus predict which product each of your prospects is likely to purchase.

The following table represents your database.

Name	Age	Residence	Socio-Occupational Category	Product Purchased
Charles	34	New Orleans	Manager/Administrator	Product A
John	37	Washington	Manager/Administrator	Product A
Marlène	31	Boston	Civil servant	Product B
Prospect 1	34	Oakland	Manager/Administrator	?
Prospect 2	24	Washington	Civil servant	?
...	...	...	...	...
Prospect n	35	Sacramento	Skilled tradesman	?

Constraints Governing Use

The following constraints govern the use of a target variable:

- Within a training data set, **all target variable values must be known**.
- Only **binary** or **continuous** variables may be used as target variables.

Explanatory Variable

Definition

An **explanatory variable** is a variable that describes your data and which serves to explain a target variable.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Synonyms

Depending upon your profile and your area of expertise, you may be more familiar with one of the following terms to refer to explanatory variables:

- Causal variables,
- Independent variables,
- Input variables.

These terms are synonyms.

Example

Your company is marketing two products A and B.

You have a database which contains references to:

- 1,500 of your customers. You know which product, A or B, each customer has purchased.
- 10,000 prospects. You want to know which product each prospect is likely to purchase.

The variables "name", "age", "address" and "socio-occupational class" are your explanatory variables: they allow you to generate a model capable of explaining and predicting the value of the target variable "product purchased".

The following table represents your database.

Name	Age	Adress	Socio-Occupational Class	Product Purchased
Charles	34	New Orleans	Manager/Administrator	Product A
John	37	Washington	Manager/Administrator	Product A
Marlene	31	Boston	Civil servant	Product B
Prospect 1	34	Oakland	Manager/Administrator	?
Prospect 2	24	Washington	Civil servant	?
...	...	...	...	...
Prospect n	35	Sacramento	Skilled tradesman	?

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Weight Variable

Definition

A **weight variable** allows one to assign a relative weight to each of the observations it describes, and actively orient the training process. To declare a variable a weight variable results in creating a number of copies of each of the data set observations, proportional to the value they possess for that variable.

Specifying a weight variable can be used:

- either to assign a higher weight to a single line
- or to do stratified sampling.

The effect of the weight can be considered as the following: a line with a weight of two in the training data set is exactly equivalent to having two identical lines with a weight of one.

Example

Imagine a data set in which the observations correspond to individual Americans. These observations are described by the variable "age", among others. Defining the variable "age" as a weight variable means that for generation of the model, older individuals will be weighted more heavily than younger individuals.

Constraints Governing Use

Only **positive continuous variables** may be used as weight variables.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.7 Models

4.7.1 Fundamental Definition

The term "model" carries many different meanings depending on its field of application. In Data Mining, a model describes and explains the relationships which exist between input data (explanatory variables) and output data (one or more target variables). It allows one to predict and explain phenomena, or to describe them.

To quote George E.P. Box, "*All Models Are Wrong But Some Are Useful*".

Note

Quote from "Robustness is the Strategy of Scientific Model Building" in Robustness in Statistics. eds., R.L. Launer and G.N. Wilkinson, 1979, Academic Press.

4.7.2 Performance of a Model

A model that has satisfactory performance is one that possesses both:

- High explanatory power, that is, sufficient capacity to explain the target variable. This explanatory power is indicated by the predictive power of the model.
- High robustness, that is, sufficient capacity to repeat the same performance on new data sets containing observations of a similar nature to the training data set. This explanatory power is indicated by the prediction confidence.

4.7.3 Types of Models

In Data Mining, there are two types of models:

- Predictive and explanatory models, which allow one to predict and explain phenomena,
- Descriptive models, which allow one to describe data sets.

With SAP InfiniteInsight®, you can generate models that are both highly descriptive and highly predictive.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.7.4 Generating a Model

SAP InfiniteInsight® models are generated during a phase called the "training phase", using a training data set.

Depending on the situation, the training data set may be cut into three sub-sets:

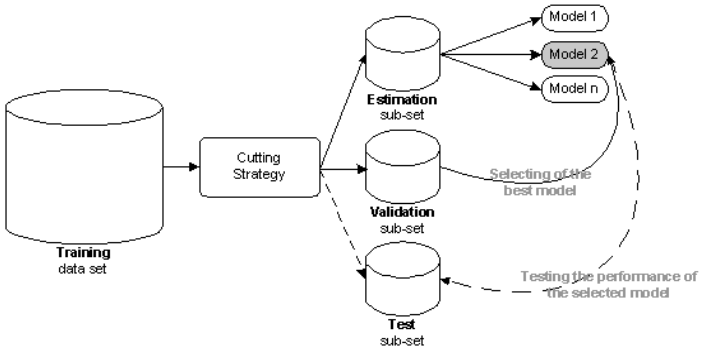
- An **Estimation** sub-set.
- A **Validation** sub-set.
- A **Test** sub-set.

A cutting strategy determines the way in which the data of the training data set are distributed across the sub-sets.

The Estimation and Validation sets are used for actual training, and the Test set (sometimes referred to as the "hold-out sample") is used to ensure that the predicted performance is correct.

> **Note**
>
> When using SAP InfiniteInsight®, the data sub-sets are virtual: they are not stored in memory at any time. The file corresponding to the initial data set remains intact at all times.

The figure below illustrates the model generation process, known as the "training phase".



4.7.5Representation of a Model

A model may be represented in many different ways, including:

- a decision tree,
- a neural network,
- a mathematical function.

In SAP InfiniteInsight®, models are represented in the form of mathematical functions, specifically, **polynomials**.

Description of the Polynomial

A polynomial may be of degree 1, 2, 3 or greater. By defining the polynomial degree, you are defining the degree of complexity of the model.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Examples of Polynomials

A polynomial of degree 1 is of the form:

```
f(X1, X2, ..., Xn) = w0 + w1.X1 + w2.X2 + ... + wn.Xn
```

A polynomial of degree 2 is of the form:

```
f(X1, X2, ..., Xn) =
w0 + w1.X1 + w2.X2 + ... + wn.Xn +
w11X1.X1 + w12.X1.X2 + w13.X1.X3 + ... + wij.Xi.Xj
```

Methodology

In the large majority of cases, a **first degree polynomial** is sufficient for generation of a relevant and robust model.

Using a higher degree of polynomial does not always guarantee better results than those obtained with a first degree polynomial. In addition, the higher the degree of polynomial you select:

- The more time needed to generate the corresponding model,
- The more time needed to apply the model to new data sets,
- The harder it is to interpret the results of modeling.

The selection of the degree of the polynomial depends on the nature of the data to be analyzed. The recommended method is to:

- First generate a model with a first order model. In the large majority of cases, this degree will be sufficient to guarantee a relevant and robust model.
- Test the results thus obtained with models of greater degree, if the performance of the first order model seems inadequate.

4.7.6 Validating the Model

Once the model has been generated, you must verify its validity by examining the performance indicators:

- The **predictive power** allows you to evaluate the **explanatory power** of the model, that is, its capacity to explain the target variable when applied to the training data set. A perfect model possesses a predictive power of 1 and a completely random model possesses a predictive power of 0.
- The **prediction confidence** defines the **degree of robustness** of the model, that is, its capacity to achieve the same explanatory power when applied to a new data set. In other words, the degree of robustness corresponds to the predictive power of the model when applied to an application data set.

To discover how these indicators are calculated, Predictive Power, Prediction Confidence and Model Graphs (on page 40).

Note

Validation of the model is a critically important phase in the overall process of Data Mining. Always be sure to assign significant importance to the values obtained for the predictive power and the prediction confidence of a model.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.7.7 Under what Circumstances is a Model Acceptable?

Quality Indicator: the Predictive Power

No minimum threshold is required for the predictive power of a model. This depends upon the context of your work, that is, your domain of application, the nature of your data and your business issue.

In some cases, a model with a predictive power as low as 0.1 may allow realization of a profit of several thousands dollars. In all cases, a positive predictive power indicates that the model generated will perform better than a random model.

Robustness Indicator: the Prediction Confidence

A model with a prediction confidence inferior to 0.95 must be considered with caution. The performance of such a model is very likely to vary between the training data set and the application data sets.

4.7.8 How to Obtain a Better Model

Obtaining a better model is achieved by:

- Improving the prediction confidence of the model, or
- Improving the predictive power of the model, or
- Improving both the predictive power and he prediction confidence of the model.

Several techniques allow you to improve these indicators:

- You can increase the degree of complexity of the model (polynomial degree).
- The following table presents other techniques.

To improve...	You can...
The **predictive power** of a model	▪ Add variables to the training data set ▪ Use combinations of explanatory variables that seem relevant to you
The prediction confidence of a model	Add observations to the training data set

Note

For more information about improving the predictive power and the prediction confidence, Indicators Specific to SAP InfiniteInsight® (see page 39).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.8 Performance Indicators

4.8.1 Indicators Specific to SAP InfiniteInsight®

Two indicators allow you to evaluate the performance of a SAP InfiniteInsight® model.

- The **predictive power**, which is the quality indicator of the model,
- The **prediction confidence**, which is the robustness indicator.

Quality Indicator: Predictive Power

Definition

The **predictive power** of a model is the quality indicator of models generated using SAP InfiniteInsight®. This indicator corresponds to the proportion of information contained in the target variable that the explanatory variables are able to explain.

Example

A model with a predictive power of:

- "0.79" is capable of explaining 79% of the information contained in the target variable using the explanatory variables contained in the data set analyzed.
- "1" is a hypothetical perfect model, capable of explaining 100% of the target variable using the explanatory variables contained in the data set analyzed. In practice, such a predictive power would generally indicate that an explanatory variable 100% correlated with the target variable was not excluded from the data set analyzed.
- "0" is a purely random model.

Improving the Predictive Power of a Model

To improve the predictive power of a model, new variables may be added to the training data set. Explanatory variables may also be combined.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Robustness Indicator: Prediction Confidence

Definition

The **prediction confidence** is the robustness indicator of the models generated using SAP InfiniteInsight®. It indicates the capacity of the model to achieve the same performance when it is applied to a new data set exhibiting the same characteristics as the training data set.

Example

A model with a prediction confidence:

- Equal to or greater than "0.98" is very robust. It has a high capacity for generalization.
- Less than "0.95" must be considered with caution. Applying it to a new data set will incur the risk of generating unreliable results.

Improving the Prediction Confidence of a Model

To improve the prediction confidence of a model, additional observation rows may be added to the training data set.

Predictive Power, Prediction Confidence and Model Graphs

On the model graph plot:

- Of the estimation data set (default plot), the **predictive power** corresponds to "the area found between the curve of the **model generated** and that of the **random model**" divided by "the area found between the curve of the **perfect model** and that of the **random model**". As the curve of the generated model approaches the curve of the perfect model, the value of the predictive power approaches 1.
- Of the estimation, validation and test data sets (select the corresponding option from the list *Data set*, located below the plot), the prediction confidence corresponds to one minus "the area found between the curve of the **estimation data set** and that of the **validation data set**" divided by "the area found between the curve of the **perfect model** and that of the **random model**".

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Advanced Users: Predictive Power for Continuous Targets

1. Working with the Validation data set, use a uniform encoding based on the distribution curve to map the target values into the range [-1;1].
   The curve is different for each sub data set. You can access to this curve through the parameter tree under the section 'UniformCurvePoints'
2. Sort the normalized target values, and graph the cumulative sums to obtain the Wizard graph. Note that 20 bins are used to give a good approximation while using less computation.
3. Re-sort by estimated values, and again graph the corresponding actual values cumulative distributions (validation graph).
4. As usual, the predictive power is the ratio of the validation and wizard areas.

The predictive power is thus based on the order of the estimates and compares this order with the order of the actual continuous targets. As such, it is more robust than the L1 (such as Mean absolute Error - MAE-) or L2 metrics (such as Mean square error -MSE-, or Root Means Square Error -RMSE-, or Pearson coefficient) often used for regression, since one very large erroneous target will never decrease the overall predictive power figure (and could be a major cause for instability of all the other metrics). On the other hand, the predictive power does not take into account the estimated values with respect to the target values: in other words, a model with estimates in the range of [-2;2] could have a very good predictive power, even if the actual targets are in the range [0;100], provided that the model has found the correct order between estimates and actual target values. InfiniteInsight® technology limits this effect by providing piece-wise linear recalibration of the estimates to the actual targets based on the statistics on the Validation data set, thus providing not only good order estimates but also good range estimates.

4.8.2 Other Commonly Used Indicators

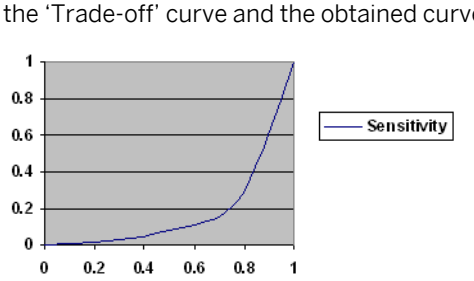
Three other indicators, commonly used in Data Mining, are provided to assess a SAP InfiniteInsight® model:

- the GINI index,
- the K-S
- the AUC

GINI Index

The Gini statistic is a measure of predictive power based on the Lorenz curve. It is proportionate to the area between the random line and the Model curve.

The GINI index is defined as the area under the Lorenz curve (see page 48). The GINI index is the area between the 'Trade-off' curve and the obtained curve multiplied by 2. This is often pictured as the following chart:



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

The horizontal axis 'grows' with the score and can be associated with 1-f.;

This is simply expressed using our notations as:

$$GINI = 2\left(\frac{1}{2} - \int_{-\infty}^{+\infty}(1-\alpha(t))d\left((1-\alpha(t)) * p_G + \beta(t) * (1-p_G)\right)\right)$$

αβ

$$GINI = 2(1-p_G)\left(\frac{1}{2} - (1-AUC)\right) = (1-p_G)(2AUC-1)$$

Using these notations, we know that the GINI index of a random model is 0 and for a perfect model is $1-p_G$.

K-S

K-S is the Kolmogorov-Smirnov statistic applied here as a measure of deviation from uniform response rates across categories of a variable. Kolmogorov-Smirnov is a non-parametric, exact goodness-of-fit statistic based on the maximum deviation between the cumulative and empirical distribution functions

In the case of a binary classification task, people are interested by the difference between the Lorenz curve (see page 48) for the good cases '1-α', and the Lorenz curve for the bad cases 'β' when selecting an increasing ratio of population. These curves evolve from 0 to 1 together, and the K-S statistics is the maximum deviation between these two curves. For a perfect system, the K-S statistics is 1 and that for a random system, because of the equality between the two curves, the K-S statistics is 0.

> **Tip**
>
> The K-S is used to calculate the difference between two distributions in order to have an idea about the quality of a data set.

AUC

The AUC statistic is a rank-based measure of model performance or predictive power calculated as the area under the Receiver Operating Characteristic curve (see "ROC" on page 47). For a simple scoring model with a binary target, this represents the observed probability of a signal (responder) observation having a higher score than a non-signal (non-responder) observation. For individual variables, ordering based on score is replaced by ordering based on the response probability for the variable's categories (for example cluster ID or age range response rates).

The corresponding equation is:

$$AUC = \int_{+\infty}^{-\infty}\alpha(t)d(1-\beta(t)) = \int_{0}^{1}\alpha(\gamma)d\gamma$$

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

So we have:

$$AUC_{perfect} = 1$$

$$AUC_{random} = \int_0^1 (\gamma) d\gamma = \left[\frac{x^2}{2} \right]_0^1 = \frac{1}{2}$$

One of the interest of the measure of AUC is its independence from the target distribution: let us imagine that we build another data set where we duplicate each 'good' example twice, the AUC of the model will be the same.

Caution

Area Under the Roc Curve (AUC) has very nice properties to evaluate a binary classification system. It is widely used now by statisticians even if it is not easy to "picture" for non-statisticians.

4.8.3 Error Indicators

First, some basic notations:

- Target (response value): y_i
- Predictor (predictor response value): $\hat{y}_i$
- Residual: $r_i = y_i - \hat{y}_i$
- Error: $u_i = | y_i - \hat{y}_i | = | r_i |$
- Weight of the tested observation: w_i
- Total weight of the population: $W = \sum_{i=1}^{n} w_i$
- Target average: $\bar{y} = \frac{1}{W} \sum_{i=1}^{N} w_i y_i$
- Predictor average: $\bar{\hat{y}} = \frac{1}{W} \sum_{i=1}^{N} w_i \hat{y}_i$

Mean Absolute Error (L1)

Definition: mean of the absolute values of the differences between predictions and actual results, (City block distance or Manhattan distance)

Formula:

$$L1 = \frac{1}{W} \sum_{i=1}^{N} w_i u_i$$

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Mean Square Error (L2)

Definition: square root of the mean of the quadratic errors (Euclidian distance or root mean squared error - RMSE)

Formula:

$$SSE_w = \sum_{i=1}^{N} w_i {u_i}^2$$

$$MSE = \frac{SSE_w}{W} = \frac{1}{W} \sum_{i=1}^{N} w_i {u_i}^2$$

$$L2 = \sqrt{MSE}$$

Maximum Error (LInf)

Definition: maximum absolute difference between predicted and actual values (upper bound) (Chebyshev distance)

Formula:

$$L\infty = \max_i u_i$$

Error Mean

Definition: mean of the difference between predictions and actual values

Formula:

- Mean Percent Error (MPE):

$$MPE = \frac{1}{W} \sum_{i=1}^{N} w_i \frac{y_i - \hat{y}_i}{y_i}$$

- Mean Absolute Percent Error (MAPE):

$$MAPE = \frac{1}{W} \sum_{i=1}^{N} w_i \frac{| y_i - \hat{y}_i |}{y_i}$$

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Error Standard Deviation

Definition: dispersion of errors around the actual result

Formula:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(u_i - \bar{u})^2}$$

where

$$\bar{u} = \frac{1}{N}\sum_{i=1}^{N}u_i$$

Classification Rate

Definition: ratio between the number of correctly classified records and the total number of records

Formula:

$$CR(t) = \frac{G\alpha(t)+B\beta(t)}{G+B} = p_G\alpha(t) + (1-p_G)\beta(t)$$

Determination Coefficient (R2)

Definition: ratio between the variability (sum of squares) of the prediction and the variability (sum of squares) of the data.

Formula:

$$SSR = \sum_{i=1}^{N} w_i(\hat{y}_i - \bar{y})^2$$

$$SST = \sum_{i=1}^{N} w_i(y_i - \bar{y})^2$$

$$R2 = \frac{SSR}{SST}$$

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4.9 Profit Type

4.9.1Definition

A **profit type** allows calculation of the profit that may be realized using the model. In general, a benefit is associated with the positive (or expected) values of the target variable and a cost is associated with the negative (or unexpected) values. For instance, in the context of a promotional mailing campaign, a person is associated with:

- A **benefit** for responding to the promotional mailing,
- A **cost** for not responding to the promotional mailing.

4.9.2 Available Profit Types

To visualize the profit that may be realized using a model generated by the SAP InfiniteInsight®, you may use the following profit types:

- **Detected** profit,
- **Lift** profit,
- **Standardized** profit,
- **Customized** profit.

Detected Profit

Detected profit is the profit type shown as the default. It allows examination of the percentage of observations belonging to the target category of the target variable, that is, the least frequent category, as a function of the proportion of observations selected from the entire data set. Using this profit:

- The value "0" is assigned to observations that do not belong to the target category of the target variable,
- The value "1" (frequency of the target category of the target variable in the data set) is assigned to observations that do belong to the target.

Lift Profit

Lift profit allows examination of the difference between a perfect model and a random model and between the model generated by SAP InfiniteInsight® and a random model.

It represents the ratio between a model and the random model, that is the performance of a model compared to a model that would only allow to select observations at random from your database.

You can thus visualize how much better your model is compared with the random model.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Standardized Profit

Standardized profit allows examination of the contribution of the model generated by SAP InfiniteInsight® features relative to a model of random type, that is, in comparison with a model that would only allow selecting observations at random from your database.

This profit is used for the plots of variable details, which present the significance of each of the categories of a given variable with respect to the target variable.

Customized Profit

Customized profit allows you to define your own profit values, that is, to associate both a cost and a benefit to each value of the target variable. For instance, you can define:

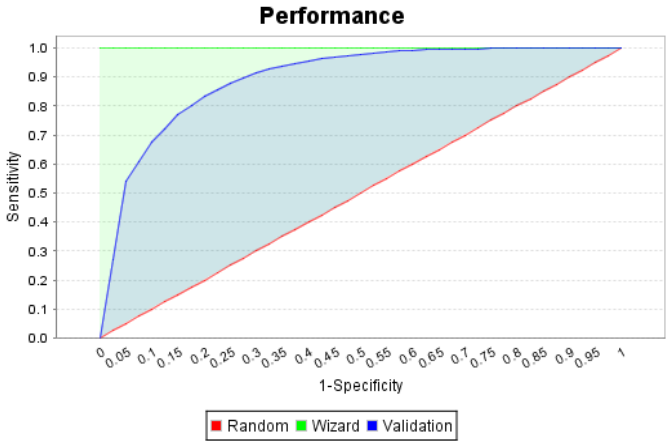
- the cost of sending out a mailing as a negative value, for example "-5".
- the benefit brought in by the response to that mailing as a positive value, for example "20".

4.10 Advanced Model Curves

In addition to the profit curves detailed in the previous section, a series of advanced model curves are available in SAP InfiniteInsight®.

4.10.1 ROC

The *ROC* (*Receiver Operating Characteristic*) graph is derived from signal detection theory. It portrays how well a model discriminates in terms of the tradeoff between sensitivity and specificity, or, in effect, between correct and mistaken detection as the detection threshold is varied.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

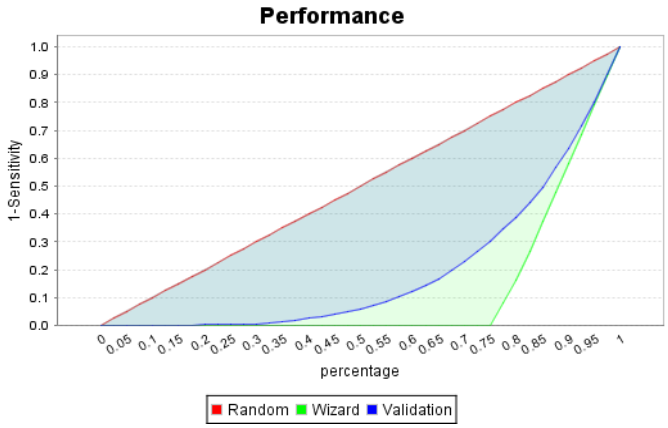
Sensitivity, which appears on the Y axis, is the proportion of CORRECTLY identified signals (true positives) found (out of all true positives on the validation data set).

[1 – Specificity], which appears on the X axis, is the proportion of INCORRECT assignments to the signal class (false positives) incurred (out of all false positives on the validation data set). (Specificity, as opposed to [1 – specificity], is the proportion of CORRECT assignments to the class of NON-SIGNALS – true negatives.)

4.10.2 Lorenz Curves

Lorenz "Good"

Lorenz 'Good' displays the cumulative proportion of missed signals (false negatives) accounted for by the records corresponding to the bottom x% of model scores.

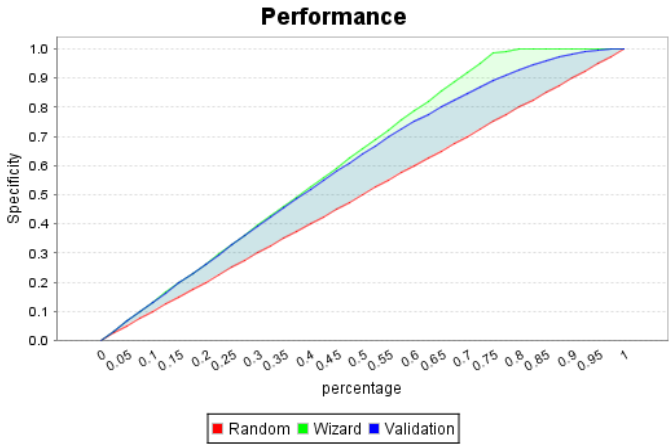


The Y axis measures *[1 – sensitivity]*, that is [1 – the proportion of true positives], which is equivalent to the proportion of missed signals or lost opportunity. Because the data are ordered from records predicted least likely to be signals on the left to records most likely to represent signals on the right, the slower the rise, the more sensitive the model in terms of detecting signals (or responders). The wizard line turns upward from the x-axis at the point corresponding to the proportion of non-signals in the validation data set.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Lorenz "Bad"

Lorenz 'Bad' displays the cumulative proportion of true negatives (specificity) accounted for by the bottom x% of model scores. Here, the faster the rise, the lower the frequency of erroneous detection.



4.10.3 Density Curves

The density curves display the density function of the variable Score in the set of Events (Curve Density "Good") and in the set of Non Events (Curve Density "Good"). These curves can also be viewed as the "derivate" of Lorenz curves (the density function is by definition the derivate of the cumulative density function).

The estimated density function in a bin or interval is equal to:

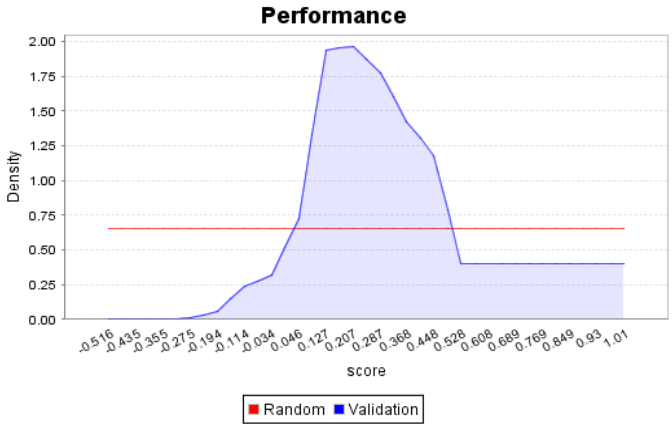
$$\frac{\frac{Number\ of\ Events\ in\ the\ Interval}{Total\ number\ of\ Events}}{Length\ of\ the\ interval}$$

The length of an interval is by definition its upper bound minus its lower bound

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

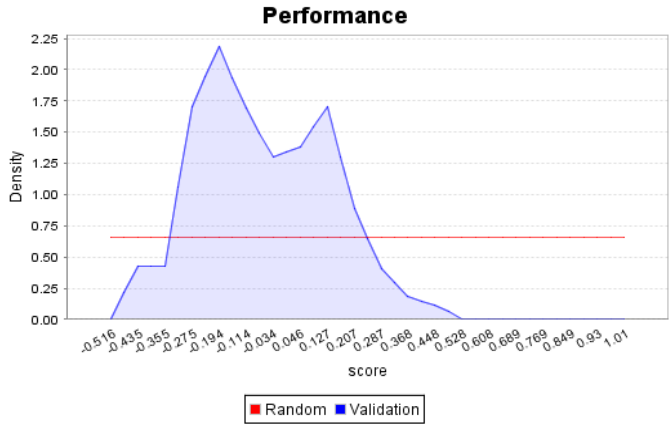
Density "Good"

This curve displays the distribution of model scores for responders/signals.



Density "Bad"

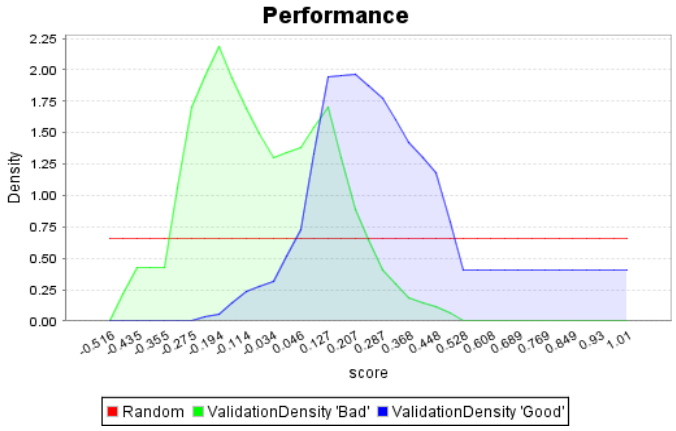
This curve displays the distribution of model scores for non-responders/non-signals.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Density "All"

This curve displays both the curves *Density "Good"* and *Density "Bad"*, thus allowing the user to compare both distributions.

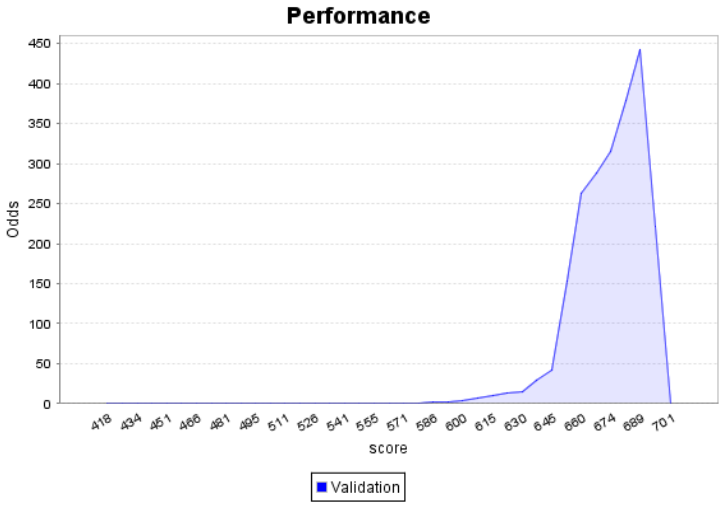


4.10.4 Risk Curves

Good/Bad Odds

The X-axis represents the risk score and the Y-axis represents the odds ratio value.

The odds ratio is equal to (1-p)/p , p is the probability of risk.

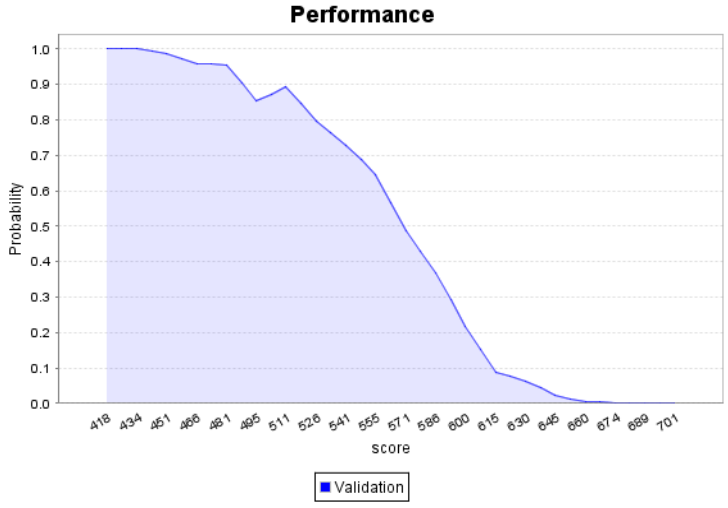


© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Probability of Risk

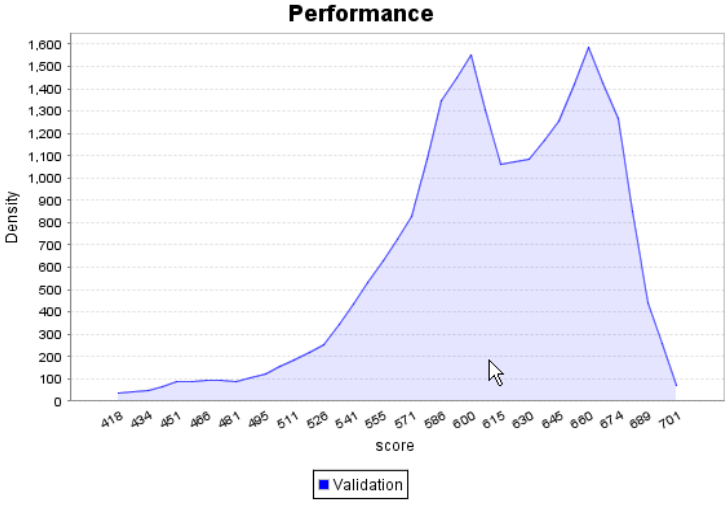
The X-axis represents the risk score and the Y-axis represents the odds ratio value.

The probability of risk p is computed for each risk score bin this way: number of "Bad" divided by the number of records in the risk score bin.



Population Density

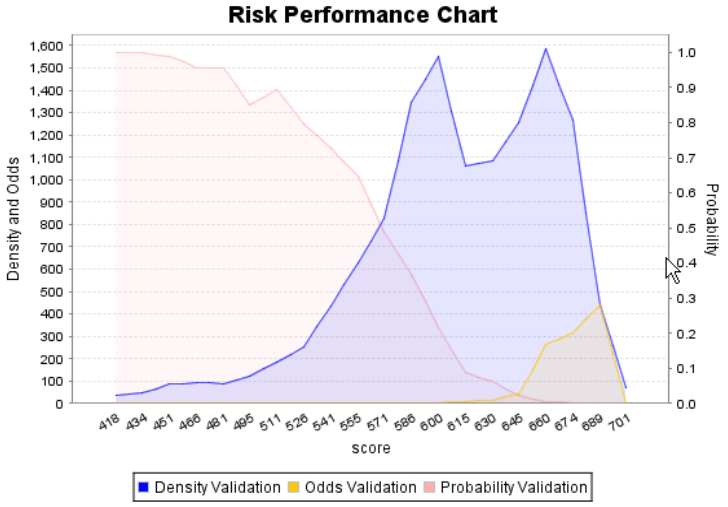
The density is computed according to the number of records in each risk score bin (20 by default).



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Risk 'All'

All three curves are displayed in the same graph. Note that the y-axis of the probability curve is on the right hand side. The y-axis of the population density and the good/bad odds on the left.



CUSTOMER
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

5 InfiniteInsight® Modeler - Regression/Classification

IN THIS CHAPTER

Application Scenario: Enhance Efficiency and Master your Budget using Modeling 54
Creating a Classification Model Using InfiniteInsight® Modeler 66

5.1 Application Scenario: Enhance Efficiency and Master your Budget using Modeling

5.1.1 Presentation

In this scenario, you are the Marketing Director of a large retail bank. The bank wants to offer a new financial product to its customers. Your project consists of launching a direct marketing campaign aimed at promoting this product. You have a large database of prospects at your disposal and a limited and closely monitored budget, and you are also subject to significant time constraints. In order to maximize the benefits of your campaign, your business issue consists of:

- **Contacting those prospects most likely to be interested in** the new financial product,
- **Identifying the ideal number of prospects to contact** out of the entire database.

Using the **InfiniteInsight® Modeler - Regression/Classification** (formerly known as K2R) feature, you can rapidly develop an explanatory and predictive model at the least possible cost. This model allows you to respond to your business issue and accomplish your objectives.

5.1.2 Your Objective

Imagine the following case.

You are the Marketing Director of a large retail bank. This bank has decided to offer its customers a **new high-end savings product**. It prepares to **launch an extensive direct marketing campaign** to promote this new product to its prospects and customers.

The bank is experiencing heavy competition and senior management, sensitive to the stakes involved in launching this new financial product, wants the marketing campaign completed **as soon as possible**.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

5.1.3 Your Means

A Limited and Closely Monitored Budget

The enterprise controls of the bank are rigorous, and the **budget** that has been allocated to you for this marketing campaign:

- Does not allow you to contact all of the bank's prospects and customers,
- May not be exceeded.

The Information at your Disposal

The Marketing Department has a database for this campaign which contains the records of 1,000,000 prospects, identified by their principal characteristics:

- Age,
- Sex,
- Employer,
- Nationality,
- Occupation,
- Education,
- Number of hours worked per week,
- ...and so on

You note that the database you have at hand is not ideal. In fact, the database contains:

- Incongruous data,
- Redundant data,
- Missing data.

Incongruent Data

The database contains alphanumeric information (such as "occupation" and "nationality") as well as numerical information (such as "age" and "unreconciled accounts").

Redundant Data

Some information in the database is redundant, such as "degree" and "education", or "degree" and "area of work".

In the field of statistics, the term "correlated variables" is used to designate such data. In classical statistical analyses, correlated variables must be processed in a particular manner. An alternate solution is to designate only one of the two correlated variables for analysis.

Since you have neither the statistical skills not the means to handle this issue of correlation between variables, you decide to leave the database as it is.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Missing Data

Some information is missing from the database. To manage this lack of information, the Information Technology department used the following convention:

- The symbol "?" means that an alphanumeric value (such as "occupation") is missing,
- The value "99999" means that a numerical value (such as "age") is missing.

Unfortunately, you have neither the time nor the necessary resources to:

- Perform a survey to fill in the missing information,
- Re-format the database.

Technical Environment

The database available to you is stored in an **RDBMS** (relational database management system) residing on a **UNIX server**, maintained by the Information Technology department of the bank. The technical constraints of this information environment are determining factors in selecting potential data analysis tools.

5.1.4 Your Approach

By virtue of the critical stakes involved in this campaign, because of your limited budget and your inability to predict customers' enthusiasm for the new product, you have chosen to minimize your risks by dividing the project into two steps:

1. **Test the marketing campaign on a sample** of 50,000 individuals extracted from the prospects database of 1,000,000 people,
2. **Global launch of the marketing campaign using the entire contents** of the prospects database.

The Test Phase of Your Marketing Campaign

The **test phase of your marketing campaign** allowed you to collect a sample of 50,000 individuals whose behavior with respect to this new product is known:

- 25% of the prospects showed themselves to be clearly interested. They chose to accept an invitation for a meeting with one of your sales channel agents,
- 75% of the prospects declined your invitation.

Your business issue consists of understanding the test results, by identifying the reasons which led certain individuals to respond favorably to your offer, and others to respond in the negative. Then you would be able to use the analytical model obtained to predict the behavior of each of the 1,000,000 prospects in your database. This would ensure that you optimized your marketing campaign by making the offer only to those individuals most likely to be interested.

The file containing the data set used for the test was sent to you by the Information Technology department of the bank in the form of a flat file (.csv). This file corresponds to the sample file *Census01.csv*, provided with SAP InfiniteInsight® and described in the **Introduction to Sample Files** (see page 60) section.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

5.1.5 Your Business Issue

Following the test phase of your campaign, your marketing database will contain:

- A list of 1,000,000 prospects,
- A list of 50,000 prospects, selected in a random manner during the test phase, and whose response to your test campaign is known. This sample, taken from your initial database, also contains missing values and correlated variables.

Your approach to the business issue consists of using the data set in its present state, as a training data set, in order to:

- **Rapidly create an explanatory and predictive model**,
- **Next, apply this model to the entire database**.

Using the model generated, you will be able to determine:

- **How many individuals contained in your prospects database you should send your mailing to**, in order to maximize the profit/return on investment of your campaign.
- How to **classify all of the individuals in your prospects database according to their « interest »** (purchasing probability) in this new product. This interest is expressed as a "score", or probability that a prospect will respond favorably to the campaign.
- **What characterizes these individuals and what is their profile?** Validate the criteria (age, socio-occupational class, degree) that explain why a person expresses interest or not in the new financial product.
- How to **simulate in real time, the likelihood of a single individual to respond favorably to a new offer**, in particular, to allow the "Call Center" of your bank or a customer service agent to immediately know the level of interest that a prospect is likely to exhibit in this financial product.
- How to record this Score in your prospects database, in order to be able to select sub-groups of the population for new campaigns at a later date.
- How to **measure the quality and reliability** (capacity of handling new individuals) **of your model**.

In order to allow you to better respond to these issues, you have access to several possible application solutions.

5.1.6 Your Solutions

To select the individuals to whom you will send a mailing, you have several possible solutions. You can use:

- A **shotgun method**,
- An **intuitive method**,
- A **classical statistical method** (for example: neural networks, Bayesian networks, logistic models, decision trees)
- The **InfiniteInsight method**.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Shotgun Method

This method consists of performing no selection on your database, and sending out a mass mailing to every person recorded in your database. This solution guarantees that all persons likely to purchase your product are contacted.

On the other hand, this runs to exorbitant costs, far exceeding your budget, and is seldom the solution applied. In addition, it runs the risk of saturating the prospects of your bank with inappropriate offers (spamming).

Intuitive Method

This method consists of performing a selection that leans on your knowledge of your field, that is to say, you send your mailing to individuals selected in an intuitive manner from your database. This solution allows you to significantly reduce the cost of your marketing campaign and make it fit your budget.

This method is not optimal, because it does not allow you to:

- Control the real costs and return on investment of your marketing operation.
- Select which prospects to contact on a basis of real returns. It is true that you probably have a relatively good understanding of which individuals stand a good chance of becoming your customers some day. But optimizing your campaign means being able to identify those prospects that have every chance of becoming customers today as a result of the current marketing campaign.
- Discover new niche prospects, that all your knowledge of the market had not previously allowed you to identify.
- Select a predefined number of individuals. Imagine that one of the constraints of your campaign consisted of contacting exactly 5,000 prospects. Your intuition may help you to select 2,400 of these. But how are you going to identify the remaining 2,600 prospects to be contacted? A purely random selection, thus completely non-optimized, might be your only solution.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Classical Statistical Method

You may decide to use a **"classical" statistical method** to better manage the effectiveness of your campaign, and thus, your budget.

On the basis of the information that you have, a Data Mining expert could create predictive models. In other words, you could ask a statistical expert to create a mathematical model that would allow you to predict the probability of a given individual to respond to your marketing campaign, as a function of his profile.

To implement this method, the statistician must:

- Perform a detailed analysis of your test campaign,
- Prepare your database down to the smallest detail, specifically, encoding the different types of data in such a way that they can be used by the analytical tools he will apply,
- Test different types of algorithms (for example: neural networks, Bayesian networks, logistic models, decision trees) and select the one best suited to your business issue.

Typically, after a few weeks, the statistician will be able to associate a value with each individual in your database, indicating the probability of being interested or not interested in your marketing campaign.

This method presents significant constraints. You must:

- Ensure that your statistical expert, perhaps from a department external to the Marketing Department, is available for the scheduled period,
- Ensure that the cost for using this scarce resource will fit into your budget,
- Spend time explaining your domain-specific business issue to him,
- Spend time understanding the results that are provided.

InfiniteInsight® Method

The simplicity and automatic nature of **SAP InfiniteInsight®** will allow you to perform the statistical analysis of your database yourself. In addition, using SAP InfiniteInsight® will allow you to obtain results in mere minutes!

SAP InfiniteInsight® uses the latest innovations of statistical sciences and also liberates the end user from the complexity of the procedures associated with statistical analysis.

Using SAP InfiniteInsight®, you will be able to create a model that allows you to:

- Determine which individuals have the highest probability (score) of being interested in your marketing campaign (predictive modeling). You may then apply the model to your entire database.
- Break out the determining factors that describe the phenomenon that you hope to model, that is, the fact of being "interested" or "not interested" in the new financial product of the bank (descriptive modeling).

The profit curve, an important validation and control tool, allows you to compare the performance of models generated using SAP InfiniteInsight® features with that of a hypothetical random model and that of a hypothetical perfect model. At the same time, it also allows you to determine the optimal number of persons to contact to maximize the profit generated by your campaign. SAP InfiniteInsight® also provides you with indicators of the quality of the model you generate (predictive power), and its capacity to generalize and remain relevant to new data sets (prediction confidence).

SAP InfiniteInsight® provides you with the means to customize your direct marketing campaign with respect to your different customer profiles, increasing your powers of persuasion.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

5.1.7 Introduction to Sample Files

This guide is accompanied by the following sample data files:

- A data file *Census01.csv*
- The corresponding description file *desc_census.csv*.

These files allow you to evaluate SAP InfiniteInsight® features and take your first steps in using it.

Census01.csv is the sample data file that you will use to follow the scenarios of **InfiniteInsight® Modeler - Regression/Classification** and **InfiniteInsight® Modeler - Segmentation/Clustering**. This file is an excerpt from the American Census Bureau database, completed in 1994 by Barry Becker.

Note

For more information about the **American Census Bureau**, see "http://www.census.gov".

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

This file presents the data on 48,842 individual Americans, of at least 17 years of age. Each individual is characterized by 15 data items. These data, or variables, are described in the following table.

Variable	Description	Example of Values
age	Age of individuals	Any numerical value greater than 17
workclass	Employer category of individuals	▪ *Private* ▪ *Self-employed-not-inc*
fnlwgt	Weight variable, allowing each individual to represent a certain percentage of the population	Any numerical value, such as "0", "2341" or 205019".
education	Level of study, represented by a schooling level, or by the title of the degree earned	▪ *11th* ▪ *Bachelors*
education-num	Number of years of study, represented by a numerical value	A numerical value between 1 and 16
marital-status	Marital status	▪ *Divorced* ▪ *Never-married*
occupation	Job classification	▪ *Sales* ▪ *Handlers-cleaners*
relationship	Position in family	▪ *Husband* ▪ *Wife*
race	Ethnicity	▪ *White* ▪ *Black*
sex	Gender	▪ *Male* ▪ *Female*
capital-gain	Annual capital gains	Any numerical value
capital-loss	Annual capital losses	Any numerical value
native country	Country of origin	▪ *United States* ▪ *France*
class	Variable indicating whether or not the salary of the individual is greater or less than $50,000	▪ "1" if the individual has a salary of greater than $50,000 ▪ "0" if the individual has a salary of less than $50,000

Note

In order to avoid complicating the InfiniteInsight® Modeler - Regression/Classification and InfiniteInsight® Modeler - Segmentation/Clustering application scenarios, the variable *fnlwgt* is used as a regular explanatory variable in these scenarios, and not as a weight variable.

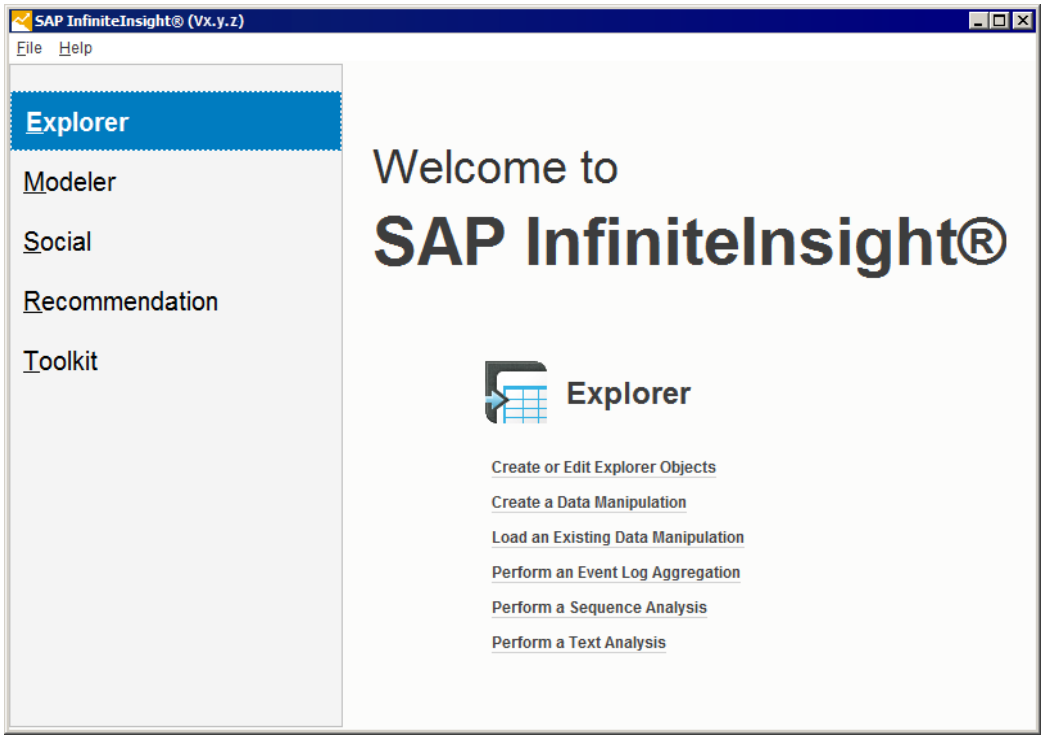
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

5.1.8 InfiniteInsight® modeling assistant

To accomplish the scenario, you will use the Java-based graphical interface of SAP InfiniteInsight®. This interface allows you to select the SAP InfiniteInsight® feature with which you will work, and help you at all stages of the modeling process.

☑ To Start InfiniteInsight® modeling assistant

1 Select *Start* > *Programs* > *SAP Business Intelligence* > *SAP SAP InfiniteInsight®* > *InfiniteInsight® modeling assistant.*
InfiniteInsight® modeling assistant screen appears.



2 Click the feature you want to use.

Editing the Options

☑ To Edit the Options of InfiniteInsight® modeling assistant

In the *File* menu, click *Preferences...*.
The window *Edit Options...* appears.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

The following options can be modified:

Category	Options
General	▪ Country ▪ Language ▪ Message Level ▪ Log Maximum Size ▪ Message Level for Strange Values ▪ Display the Parameter Tree ▪ Number of Store in the History ▪ Always Exit without Prompt ▪ Include Test in Default Cutting Strategy
Stores	▪ Default Store for Apply-in Data Set ▪ Default Store for Apply-out Data Set ▪ Default Store to Save Models
Metadata Repository	▪ Enable Single **Metadata Repository** ▪ Edit Variable Pool Content
Graphic	▪ Profit Curve Points ▪ Bar Count Displayed ▪ No InfiniteInsight® Lo*OK* and Feel ▪ Display 3D Chart ▪ Disable Double Buffering ▪ Optimize for Remote Display ▪ Remember Size and Position when Leaving
Report	▪ Number of Variables of Interest ▪ Active Style Sheet ▪ Customize Style Sheets

Customizing Style Sheets

SAP InfiniteInsight® offers the possibility to customize the generated reports. The default style sheet, called *SAP InfiniteInsight® Report Style Sheet (default)*, cannot be modified. You have to create your own style sheets to modify the settings.

Note

To create, load or save a style sheet, you have to indicate a data source in the panel *Edit Options...* before opening the window *SAP InfiniteInsight® Report Style Sheet Editor*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Create a New Style Sheet

1. In the field *Folder*, click the button (*Browse*).
2. Select a folder.
   This folder is your style sheets repository.
3. Click the button (*Add*).
   A new style sheet has been created.
4. Click the button .
   The panel *Report Style Sheet Editor* opens.
5. In the field *Style Sheet Name*, enter a name for the new style sheet.
   The extension `.krs` is automatically added.

Note

You can duplicate a style sheet by changing the name of your style sheet. The previous one is not deleted.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Delete a Style Sheet

1. Select one of the displayed style sheets.
2. Click the button (*Remove*). Note that the style sheet is not only deleted from the list but also from the data source.

☑ To Edit the General Settings

Settings...	Options...	Note...
Reports Background Color	▪ choose a color ▪ make transparent	Only the PDF and HTML formats can display a background color.
Edit Configuration	▪ font size ▪ font style ▪ font color ▪ text background color ▪ table configuration	Check the option *Dynamically render option changes* or click *Apply* when editing the settings, so that you can visualize the result.

The selected settings will be applied to both the wizard and the generated reports.

☑ To Edit the Charts Settings

Settings...	Options...	Note...
Chart Colors	▪ modify the charts colors	
Default Chart Bars Orientation	▪ horizontal ▪ vertical	It is possible to set another default orientation for specific report items.

☑ To Edit Report Items

1. Set the properties of your choice.
2. Click *Save* to validate.
   A window opens, indicating that your style sheet has been successfully saved.
3. Click *OK*.

Properties...	Functions...
Displayed as	name of the label
View Type	choose between Tabular, HTML and Graphical. The last one is only available if the report item can be displayed as a graph.
Chart Type	select one of the proposed chart types Note that this option is only available for report items of the view type *Graphical*.
Switch Bar Orientation	this option allows having another bar orientation as the default one for a specific report item
Sort by/Sort Order	you can select a column to sort by and choose between an ascending or a descending order
Visibility	you can hide columns of a report item or even menu items Note that at least one column of a report item must remain visible.

☑ To Apply the New Style Sheet to the Generated Reports

4. In the panel *Report*, select the new style sheet.
5. Click *OK*.
   A window opens, indicating that you have to restart the modeling assistant to take the edited options into account.
6. Click *OK*.
   When training a model, all the generated reports (the learn/excel/statistical reports) are now customized.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Defining a Metadata Repository

The metadata repository allows you to specify the location where the metadata should be stored.

☑ To Define a Metadata Repository

1. Choose between storing the metadata in the same place as the data or in a single place by checking the option of your choice.
2. In the list *Data Type*, select the type of data you want to access. For some type of data, you will need a specific license.
3. Use the *Browse* button corresponding to the *Folder* field to select the folder or database containing the data. In case of a protected database, you will need to enter the user name and the password in the fields *User* and *Password*.
4. Click the button *Edit Variable Pool Content* to edit the parameters of the variables stored in the variable pool.
5. Click *OK* to validate.

5.2 Creating a Classification Model Using InfiniteInsight® Modeler

Data modeling with InfiniteInsight® Modeler - Regression/Classification is subdivided into four broadly defined stages:

1. **Defining** the Modeling Parameters
2. **Generation and Validation** of the Model
3. **Analysis and Understanding** of the Analytical Results
4. **Using** a Generated Model

5.2.1 Step 1 - Defining the Modeling Parameters

To respond to your business issue you want to:

- **Identify and understand** the factors that determine whether a prospect reacts positively or negatively to your marketing campaign.
- Thereby be able to **predict** the behavior of new prospects with respect to your campaign.

The InfiniteInsight® Modeler - Regression/Classification feature (formerly known as K2R) allows you to create explanatory and predictive models.

The first step in the modeling process consists of defining the modeling parameters:

1. Select a **data source** (on page 67) to be used as training data set.
2. **Describe the data set** (on page 70) selected.
3. Select the **variables**: the **target variables**, the **explanatory variables** and possibly a **weight variable**.
4. Check the **modeling parameters**.
5. Setting **the Advanced Parameters** (on page 99) (degree, target key, rule mode, variable auto-selection and correlations). This step is optional.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Selecting a Data Source

For this Scenario

Use the file *Census01.csv* as a training data set.

This file represents the sample that you had extracted from your database and used for the test phase of your direct marketing campaign. As specified in your test plan, this file contains data concerning 50,000 prospects, for whom you now know the behavior with respect to the new financial product:

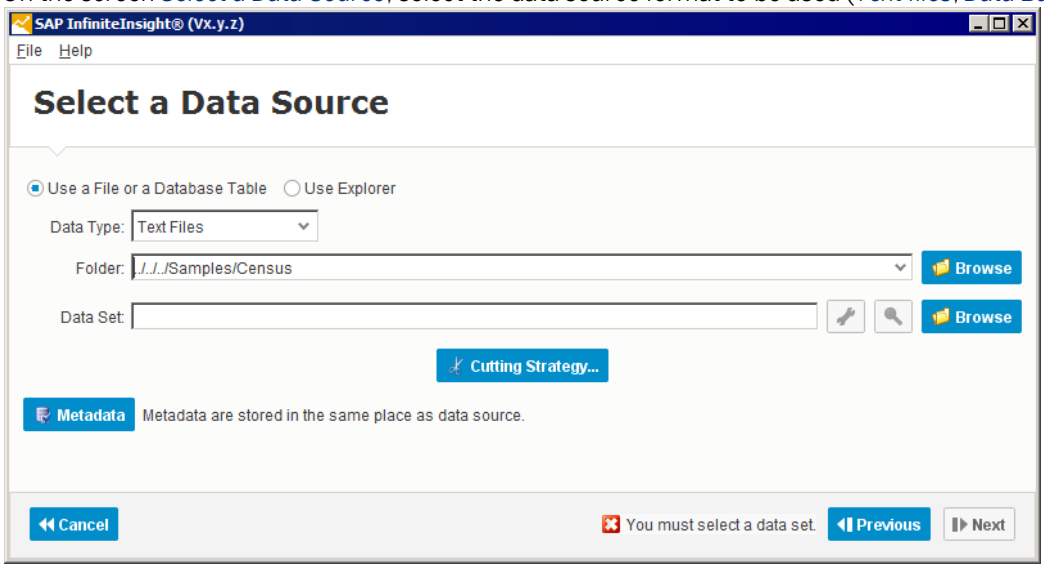
- 25% of the prospects showed themselves to be clearly interested. They chose to accept a an invitation for a meeting with one of your sales channel agents,
- 75% of the prospects declined your invitation.

In this file, you created a new variable Class, which corresponds to the reaction of prospects contacted during the test. You assigned:

- The value "1" to those prospects who responded positively to your invitation,
- The value "0" to those prospects who responded negatively to your invitation.

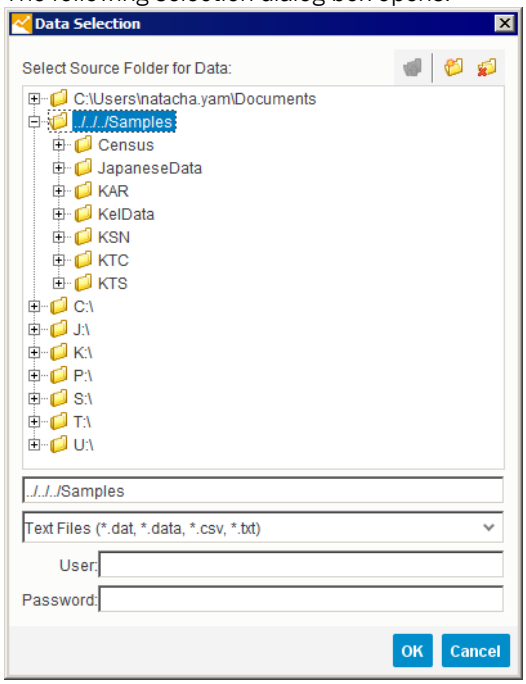
To Select a Data Source

1 On the screen *Select a Data Source*, select the data source format to be used (*Text files*, *Data Base*, ...).



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2 Click the *Browse* button.
The following selection dialog box opens.

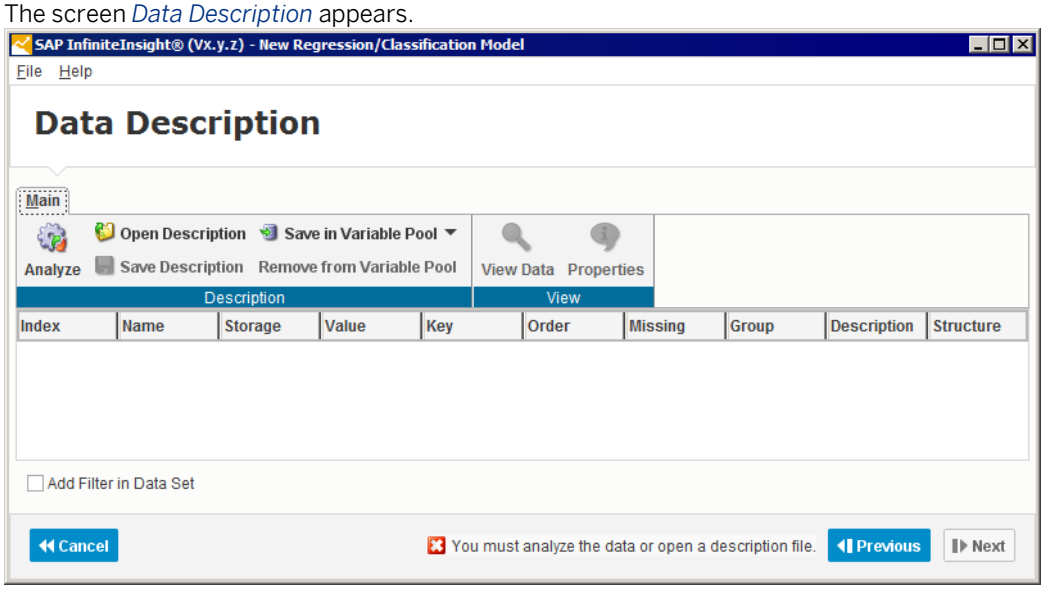


3 Double-click the *Samples* folder, then the *Census* folder.

> **Note**
>
> Depending on your environment, the *Samples* folder may or may not appear directly at the root of the list of folders. If you selected the default settings during the installation process, you will find the *Samples* folder located in *C:\Program Files\SAP InfiniteInsight\InfiniteInsightVx.y.z*

1 Select the file *Census01.csv*, then click *OK*.
The name of the file appears in the *Estimation* field.

2 Click *Next.*
The screen *Data Description* appears.



3 Go to section **Describing the Data Selected**.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Special Case: Data Stored in Databases - the Explain Mode

Before requesting data stored in a Teradata(1), Oracle(2) or SQLServer 2005 database, SAP InfiniteInsight® uses a feature, called the **Explain mode**, which categorizes the performances of SQL queries in several classes defined by the user. In order to be as fast and as light as possible, this categorization is done **without actually executing the full SQL query**.

> **Note**
>
> (1) For all versions of Teradata.
> (2) For all versions above and including Oracle 10.

The objective is to allow estimating the workload of the SQL query before executing it and then deciding --possibly thanks to an IT Corporate Policy-- if the SQL query can actually be used.

For example, an IT Corporate Policy may favor interactivity and then define 3 classes of SQL queries, each with its maximum time:

- *Immediate*: *duration < 1 s*. The query is accepted and executed immediately.
- *Batched*: *1s <= duration < 2 s*. The query is accepted but will be executed on next idle time.
- *Rejected*: *2s <= duration*. The query will never be executed.

The number, names and limits of classes are defined by the user in order for these values to match the current DBMS configuration and DBMS usage policy.

The Explain Mode has been Configured

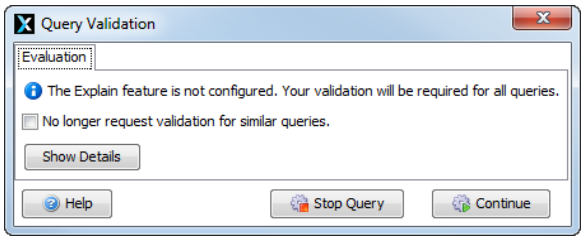
If the Explain mode has been configured by your DBMS administrator, there are two possible outcomes to a query:

- **the query is accepted and executed**: this is completely transparent. SAP InfiniteInsight® accesses the data without further input from the user.
- **the query needs to be validated before being executed**: a pop-up window opens displaying a message configured by the DBMS administrator. A query that needs validation can be categorized in two ways:
  - **medium-sized**
    You will probably have to check with your administrator to know which action to take:
- If the administrator authorizes the query, click the *Continue* button. The pop-up window closes and the requested action is carried out.
- If the administrator does not authorize the query, click the button *Stop Query*, the pop-up window closes, but no action is executed.
  - **huge**
    It means that the query will take too much time and resources. In that case, the behavior of the *Continue* button depends on the configuration set by the DBMS Administrator (for example, it can automatically refuse queries that are considered too heavy). In any case, you should check with them to know the line of action to follow.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

The Explain Mode has not been Configured

If your DBMS Administrator has not configured the Explain mode, the following pop-up opens when you try to access the data:



You need to contact your Administrator who will tell you which action to take and configure the Explain mode.

If the Administrator validates the execution of the query, you may want all queries with the same duration to be executed without validation. In that case, check the box *Do not request validation anymore for similar requests*. The validation message will then only appear for larger queries. This configuration will only be used for the current session, when closing SAP InfiniteInsight®, it will be lost. For a permanent configuration, see your DBMS Administrator, who will find the necessary information in the support document Explain Mode available in section Support and Integration Documentation of SAP InfiniteInsight® documentation.

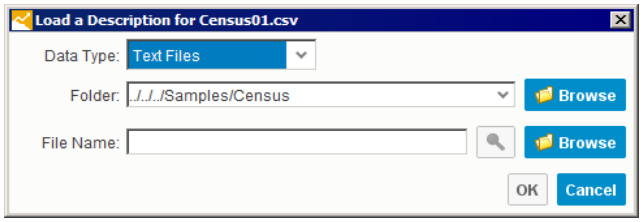
Describing the Data Selected

For this scenario

- Select *Text Files* as the file type.
- Use the file *Desc_Census01.csv* as the description file for the *Census01.cvs* data file.

To Select a Description File

1 On the screen *Data Description*, click the button *Open Description*. The following window opens:



2 In the window *Load a Description*, select the type of your description file.

3 In the *Folder* field, select the folder where the description file is located with the *Browse* button.

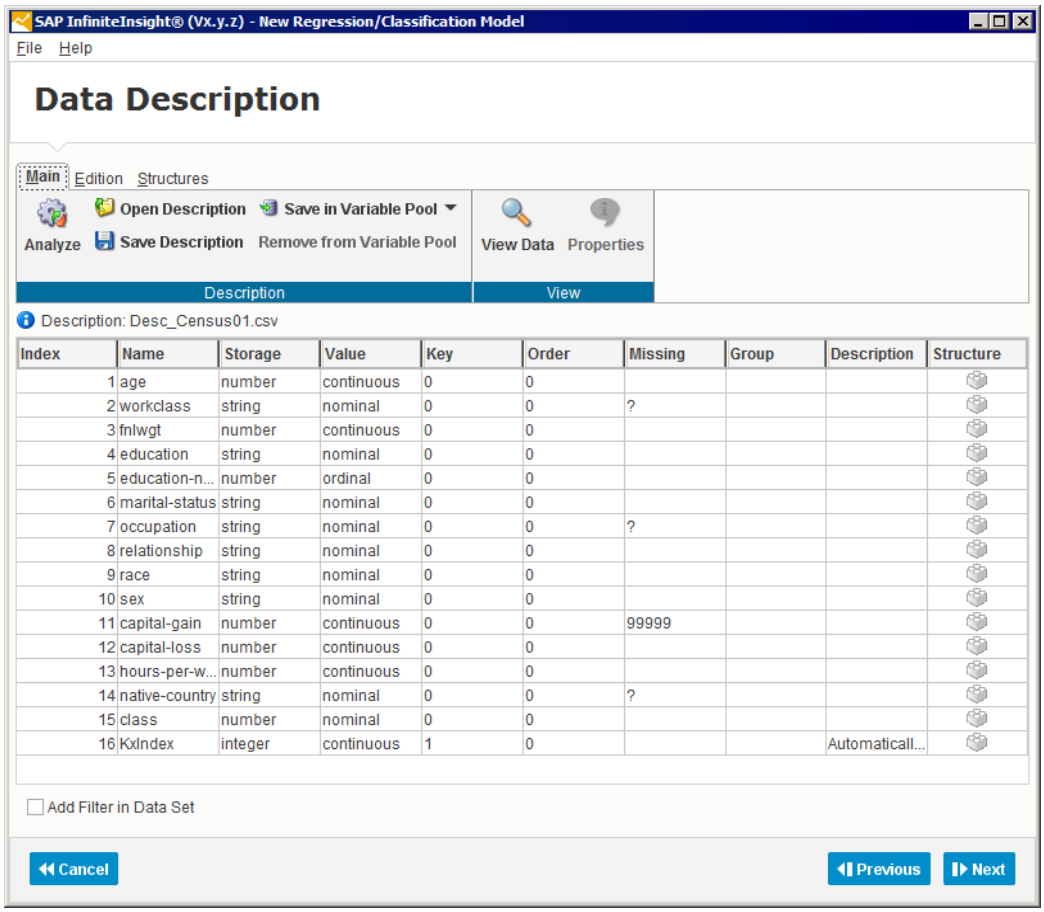
> **Note**
>
> The folder selected by default is the same as the one you selected on the screen *Data to be Modeled*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4 In the *Description* field, select the file containing the data set description with the *Browse* button.

> **Caution**
>
> When the space used for model training **contains** a physical variable named `KxIndex`, it is not possible to use a description file without any key for the described space.
> When the space used for model training **does not contain** a physical variable named `KxIndex`, it is not possible to use a description file including a description about a KxIndex variable since it does not exist in current space.

5 Click *OK*. The window *Load a Description* closes and the description is displayed on the screen *Data Description*.



Index	Name	Storage	Value	Key	Order	Missing	Group	Description	Structure
1	age	number	continuous	0	0				
2	workclass	string	nominal	0	0	?			
3	fnlwgt	number	continuous	0	0				
4	education	string	nominal	0	0				
5	education-n...	number	ordinal	0	0				
6	marital-status	string	nominal	0	0				
7	occupation	string	nominal	0	0	?			
8	relationship	string	nominal	0	0				
9	race	string	nominal	0	0				
10	sex	string	nominal	0	0				
11	capital-gain	number	continuous	0	0	99999			
12	capital-loss	number	continuous	0	0				
13	hours-per-w...	number	continuous	0	0				
14	native-country	string	nominal	0	0	?			
15	class	number	nominal	0	0				
16	KxIndex	integer	continuous	1	0			Automaticall...	

6 Click *Next*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Why Describe the Data Selected?

In order for the SAP InfiniteInsight® features to interpret and analyze your data, the data must be described. To put it another way, the description file must specify the nature of each variable, determining their:

- **Storage** format: number (*number*), character string (*string*), date and time (*datetime*) or date (*date*).

  ### Note

  When a variable is declared as *date* or *datetime*, the date coder feature automatically extracts date information from this variable such as the day of the month, the year, the quarter and so on. Additionnal variables containing this information are created during the model generation and are used as input variables for the model.

  The date coder feature is disabled for Time Series.

- **Type:** *continuous*, *nominal*, *ordinal* or *textual*.

For more information about data description, see **Types of Variables** (on page 26) and **Storage Formats (on page 29)** .

How to Describe Selected Variables

To describe your data, you can:

- Either **use an existing description file**, that is, taken from your information system or saved from a previous use of SAP InfiniteInsight® features,
- Or **create a description file using the** *Analyze* **option**, available to you in InfiniteInsight® modeling assistant. In this case, **it is important that you validate the description file obtained**. You can save this file for later re-use. If you name the description file **KxDoc_<SourceFileName>**, it will be automatically loaded when clicking the *Analyze* button.

  ### Caution

  The description file obtained using the *Analyze* option results from the analysis of the first 100 lines of the initial data file. In order to avoid all bias, we encourage you to mix up your data set before performing this analysis.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Each variable is described by the fields detailed in the following table:

The Field...	Gives information on...
Name	the variable name (which cannot be modified)
Storage	the type of values stored in this variable: ▪ *Number*: the variable contains only "computable" numbers (be careful a telephone number, or an account number should not be considered numbers) ▪ *String*: the variable contains character strings ▪ *Datetime*: the variable contains date and time stamps ▪ *Date*: the variable contains dates
Value	the value type of the variable: ▪ *Continuous*: a numeric variable from which mean, variance, etc. can be computed ▪ *Nominal*: categorical variable which is the only possible value for a string ▪ *Ordinal*: discrete numeric variable where the relative order is important ▪ *Textual*: textual variable containing phrases, sentences or complete texts **Caution** When creating a text coding model, if there is not at least one textual variable, you will not be able to go to the next panel.
Key	whether this variable is the key variable or identifier for the record: ▪ *0* the variable is not an identifier; ▪ *1* primary identifier; ▪ *2* secondary identifier...
Order	whether this variable represents a natural order. (0: the variable does not represent a natural order; 1:the variable represents a natural order). If the value is set at 1, the variable is used in SQL expressions in an "`order by`" condition. There must be at least one variable set as Order in the Event data source. **Caution** If the data source is a file and the variable stated as a natural order is not actually ordered, an error message will be displayed before model checking or model generation.
Missing	the string used in the data description file to represent missing values (e.g. "999" or "#Empty" - without the quotes)
Group	the name of the group to which the variable belongs. Variables of a same group convey a same information and thus are not crossed when the model has an order of complexity over 1 . This parameter will be usable in future version.
Description	an additional description label for the variable
Structure	this option allows you to define your own variable structure, which means to define the variables categories grouping.

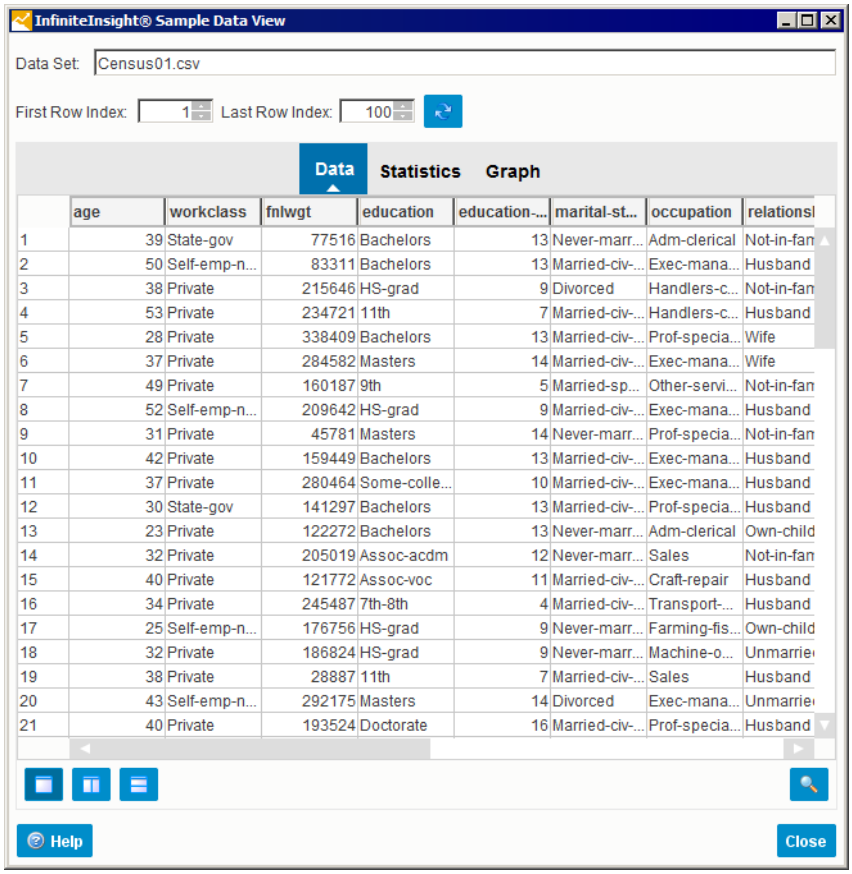
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Viewing the Data

To help you validate the description when using the *Analyze* option, you can display the first hundred lines of your data set.

☑ To View the Data

1 Click the button *View Data*. A new window opens displaying the data set top lines:



	age	workclass	fnlwgt	education	education-...	marital-st...	occupation	relationsl
1	39	State-gov	77516	Bachelors	13	Never-marr...	Adm-clerical	Not-in-fan
2	50	Self-emp-n...	83311	Bachelors	13	Married-civ-...	Exec-mana...	Husband
3	38	Private	215646	HS-grad	9	Divorced	Handlers-c...	Not-in-fan
4	53	Private	234721	11th	7	Married-civ-...	Handlers-c...	Husband
5	28	Private	338409	Bachelors	13	Married-civ-...	Prof-specia...	Wife
6	37	Private	284582	Masters	14	Married-civ-...	Exec-mana...	Wife
7	49	Private	160187	9th	5	Married-sp...	Other-servi...	Not-in-fan
8	52	Self-emp-n...	209642	HS-grad	9	Married-civ-...	Exec-mana...	Husband
9	31	Private	45781	Masters	14	Never-marr...	Prof-specia...	Not-in-fan
10	42	Private	159449	Bachelors	13	Married-civ-...	Exec-mana...	Husband
11	37	Private	280464	Some-colle...	10	Married-civ-...	Exec-mana...	Husband
12	30	State-gov	141297	Bachelors	13	Married-civ-...	Prof-specia...	Husband
13	23	Private	122272	Bachelors	13	Never-marr...	Adm-clerical	Own-child
14	32	Private	205019	Assoc-acdm	12	Never-marr...	Sales	Not-in-fan
15	40	Private	121772	Assoc-voc	11	Married-civ-...	Craft-repair	Husband
16	34	Private	245487	7th-8th	4	Married-civ-...	Transport-...	Husband
17	25	Self-emp-n...	176756	HS-grad	9	Never-marr...	Farming-fis...	Own-child
18	32	Private	186824	HS-grad	9	Never-marr...	Machine-o...	Unmarrie
19	38	Private	28887	11th	7	Married-civ-...	Sales	Husband
20	43	Self-emp-n...	292175	Masters	14	Divorced	Exec-mana...	Unmarrie
21	40	Private	193524	Doctorate	16	Married-civ-...	Prof-specia...	Husband

2 In the field *First Row Index*, enter the number of the first row you want to display.

3 In the field *Last Row Index*, enter the number of the last row you want to display.

4 Click the *Refresh* button to see the selected rows.

A Comment about Database Keys

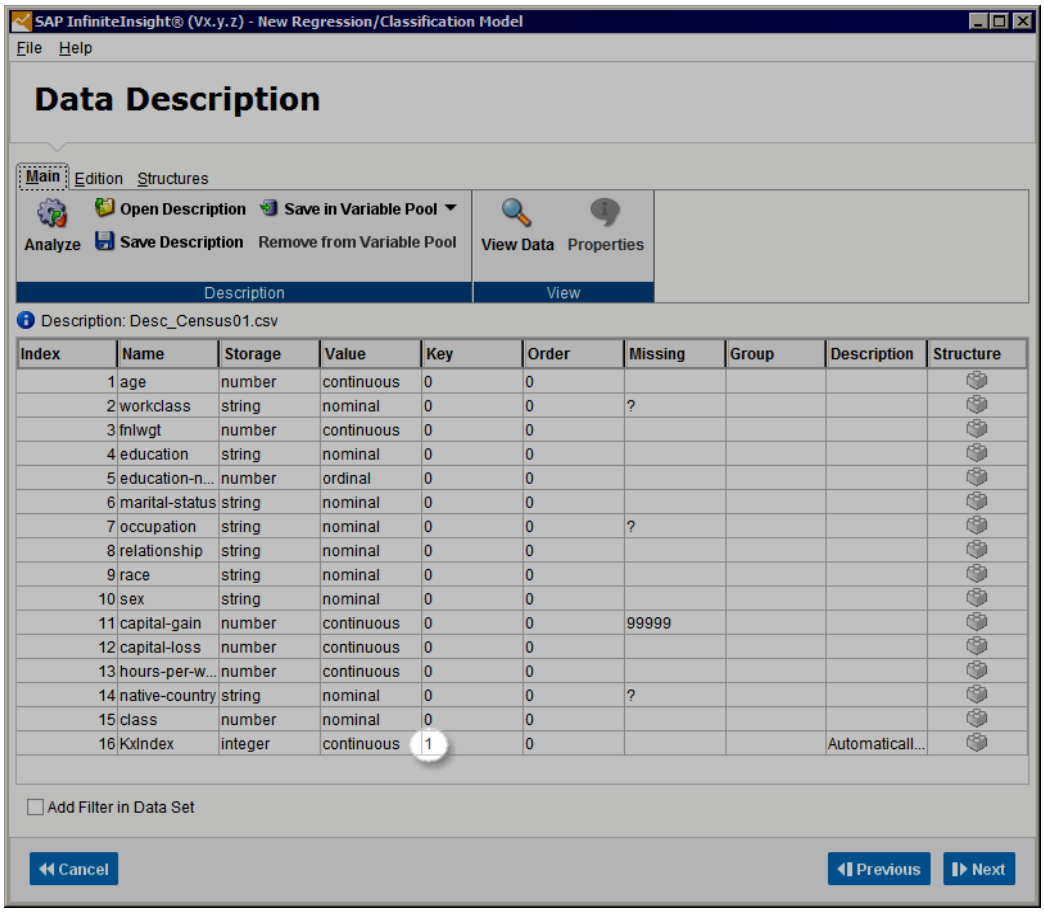
For data and performance management purposes, the data set to be analyzed must contain a variable that serves as a key variable. Two cases should be considered:

- If the initial data set **does not contain a key variable**, a variable index *KxIndex* is automatically generated by SAP InfiniteInsight® features. This will correspond to the row number of the processed data.
- If the file **contains one or more key variables**, they are not recognized automatically. You must specify them manually in the data description. See the procedure To Specify that a Variable is a Key (see page 75).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

To Specify that a Variable is a Key

1 In the *Key* column, click the box corresponding to the row of the key variable.
2 Type in the value "*1*" to define this as a key variable.



Index	Name	Storage	Value	Key	Order	Missing	Group	Description	Structure
1	age	number	continuous	0	0				
2	workclass	string	nominal	0	0	?			
3	fnlwgt	number	continuous	0	0				
4	education	string	nominal	0	0				
5	education-n...	number	ordinal	0	0				
6	marital-status	string	nominal	0	0				
7	occupation	string	nominal	0	0	?			
8	relationship	string	nominal	0	0				
9	race	string	nominal	0	0				
10	sex	string	nominal	0	0				
11	capital-gain	number	continuous	0	0	99999			
12	capital-loss	number	continuous	0	0				
13	hours-per-w...	number	continuous	0	0				
14	native-country	string	nominal	0	0	?			
15	class	number	nominal	0	0				
16	KxIndex	integer	continuous	1	0			Automaticall...	

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Defining a Variable Structure

There are three ways to define a variable structure:

- by first extracting the categories from the variable statistics, then editing or validating the suggested structure.
- by importing the structure from an existing model.
- by building a new structure from scratch.

The option *Optimal Grouping* allows you to let Data Encoding group together the categories groups defined in the variable structure if they bring the same information.

The last column of the description table indicates the state of the structure of each variable. The following table lists the possible states of a variable structure.

Icon	State	Description
	undefined	Data Encoding will automatically determine the categories grouping depending on their interaction with the target variable.
	non-editable	The structure for an ordinal string variable cannot be modified.
	defined by extraction from the variable statistics	The user must open and validate the variable structure.
	defined by the user or imported from an existing model	

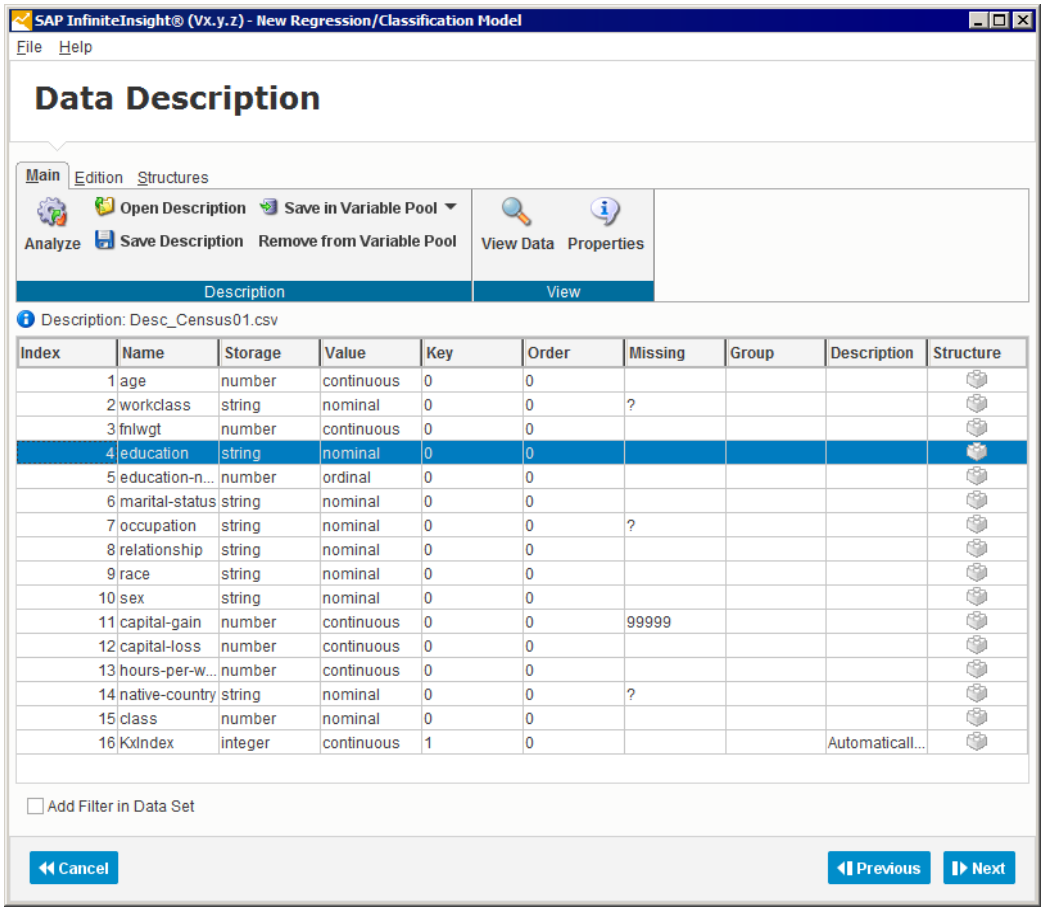
Note

A translation of the variable categories has no influence on the variable structure, which has to be set according to the original values of the variable.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Extract a Variable Structure

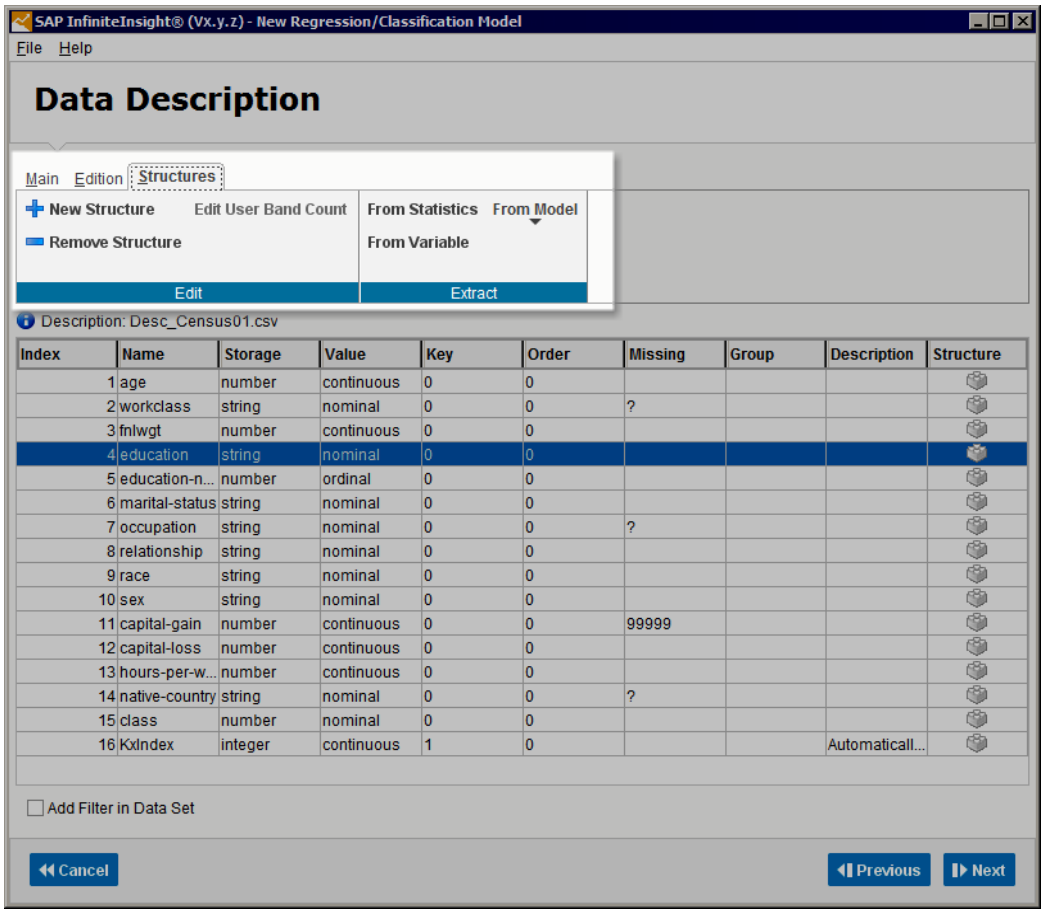
1 Select the variables for which you want to extract the structure.



Index	Name	Storage	Value	Key	Order	Missing	Group	Description	Structure
1	age	number	continuous	0	0				
2	workclass	string	nominal	0	0	?			
3	fnlwgt	number	continuous	0	0				
4	education	string	nominal	0	0				
5	education-n...	number	ordinal	0	0				
6	marital-status	string	nominal	0	0				
7	occupation	string	nominal	0	0	?			
8	relationship	string	nominal	0	0				
9	race	string	nominal	0	0				
10	sex	string	nominal	0	0				
11	capital-gain	number	continuous	0	0	99999			
12	capital-loss	number	continuous	0	0				
13	hours-per-w...	number	continuous	0	0				
14	native-country	string	nominal	0	0	?			
15	class	number	nominal	0	0				
16	KxIndex	integer	continuous	1	0			Automaticall...	

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2 Go to the *Structures* tab of the ribbon, the available options are separated in two parts: *Edit* and *Extract*.


SAP InfiniteInsight® (Vx.y.z) - New Regression/Classification Model
File Help
Data Description
Main Edition Structures
New Structure
Edit User Band Count
Remove Structure
From Statistics
From Model
From Variable
Edit
Extract
Description: Desc_Census01.csv
Index Name Storage Value Key Order Missing Group Description Structure
1 age number continuous 0 0
2 workclass string nominal 0 0 ?
3 fnlwgt number continuous 0 0
4 education string nominal 0 0
5 education-n... number ordinal 0 0
6 marital-status string nominal 0 0
7 occupation string nominal 0 0 ?
8 relationship string nominal 0 0
9 race string nominal 0 0
10 sex string nominal 0 0
11 capital-gain number continuous 0 0 99999
12 capital-loss number continuous 0 0
13 hours-per-w... number continuous 0 0
14 native-country string nominal 0 0 ?
15 class number nominal 0 0
16 KxIndex integer continuous 1 0 Automaticall...
Add Filter in Data Set
Cancel
Previous
Next

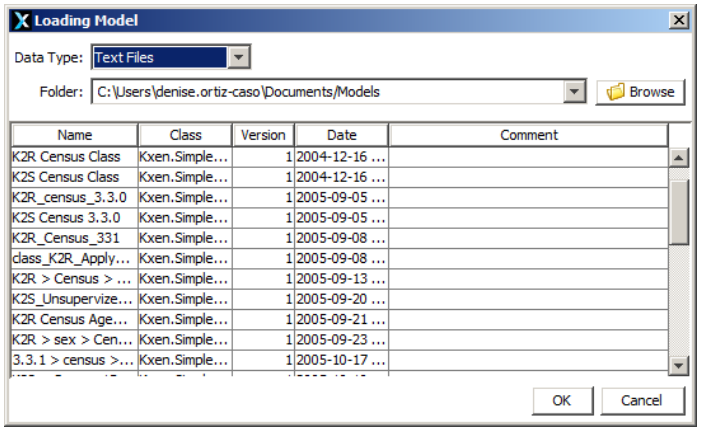
3 Click the option *From Statistics* in the *Extract* part. A progression bar is displayed while the structure is being extracted. Once the extraction is done, the icons corresponding to the selected variables change, indicating that the operation was a success and allowing you to easily identify them. You can then modify the variable structure as you need.

☑ To Import the Variable Structure from a Model

1 Select the variables for which you want to extract the structure.

2 Go to the *Structures* tab of the ribbon, the available options are separated in two parts: *Edit* and *Extract*.

3 Select the option *From Model* and choose the desired option. The window *Loading Model* opens.


Loading Model
Data Type: Text Files
Folder: C:\Users\denise.ortiz-caso\Documents\Models
Browse
Name Class Version Date Comment
K2R Census Class Kxen.Simple... 1 2004-12-16 ...
K2S Census Class Kxen.Simple... 1 2004-12-16 ...
K2R_census_3.3.0 Kxen.Simple... 1 2005-09-05 ...
K2S Census 3.3.0 Kxen.Simple... 1 2005-09-05 ...
K2R_Census_331 Kxen.Simple... 1 2005-09-08 ...
class_K2R_Apply... Kxen.Simple... 1 2005-09-08 ...
K2R > Census > ... Kxen.Simple... 1 2005-09-13 ...
K2S_Unsupervize... Kxen.Simple... 1 2005-09-20 ...
K2R Census Age... Kxen.Simple... 1 2005-09-21 ...
K2R > sex > Cen... Kxen.Simple... 1 2005-09-23 ...
3.3.1 > census >... Kxen.Simple... 1 2005-10-17 ...
OK
Cancel

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

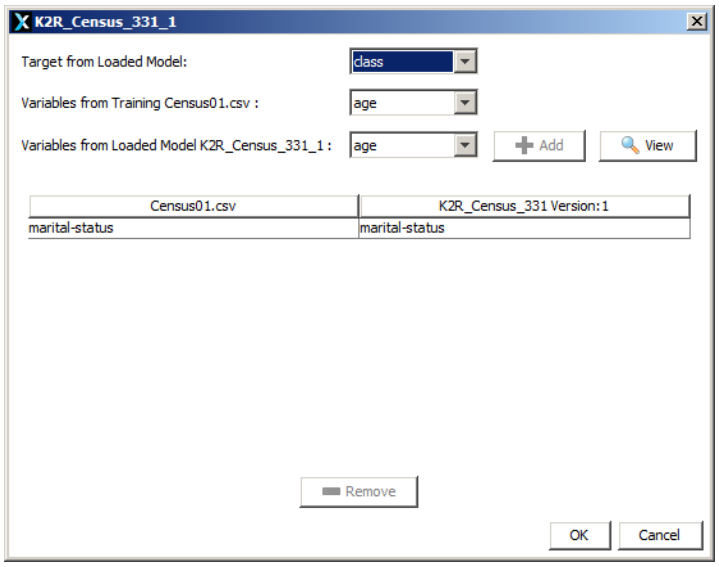
4 In the list *Data Type*, select the type of store the model is saved in.

5 Use the *Browse* button located next to the *Folder* field to select the folder or database containing the model.

6 In the displayed models list, select the model from which you want to extract the variable structure.

7 Click *OK*.

8 In the list *Target from Loaded Model*, select the target of the model. The variables you have selected are displayed in a list with the corresponding variables from the loaded model.



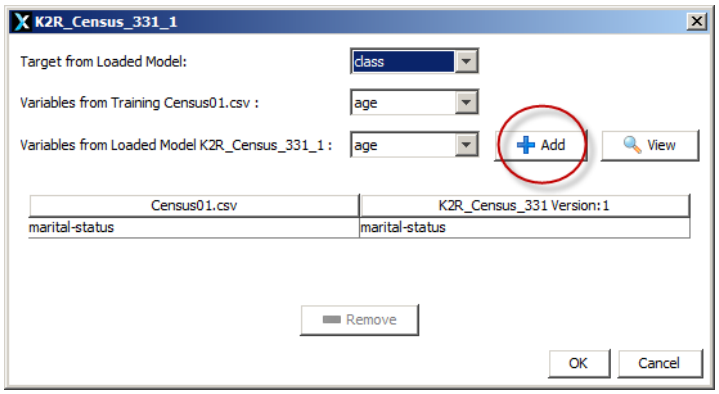
You can add or remove variables from this list and view the model variables structure as explained below.

9 Once all the variables for which you want to import the structure from the model are displayed in the list, Click *OK*. The selection window closes and the structure state changes.

☑ To Add a Variable to the List of Variables

1 In the list *Variable from Loaded Model*, select the variable you want to add to the list of variables for which the structure will be imported.

2 Click the *Add* button.



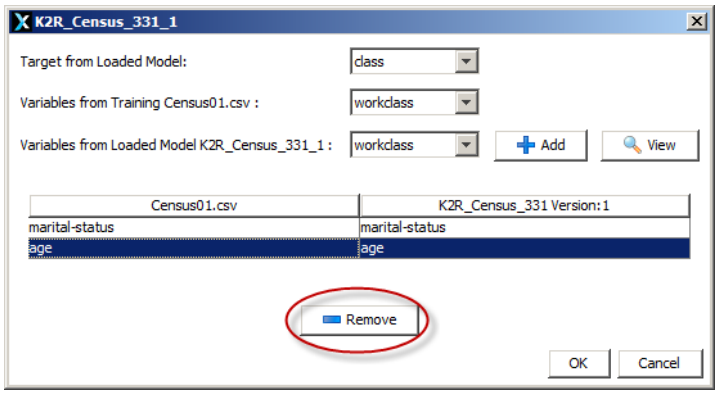
The variable appears in the list below.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Remove a Variable from the List of Variables

1 In the list located in the lower part of the panel, select the variable for which you do not want to import the structure.

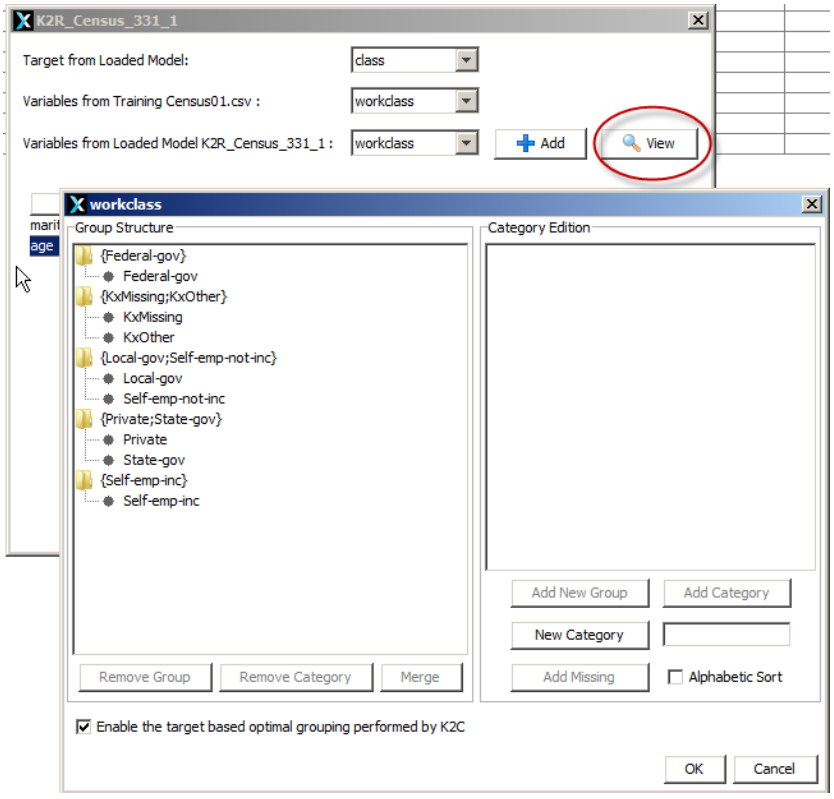
2 Click the *Remove* button.



The variable is removed from this list and added to the list *Variable from Loaded Model*.

☑ To View a Variable Structure Defined in the Loaded Model

- If the variable has not been added yet to the list of variables located on the lower part of the panel:
  1. In the list *Variable from Loaded Model*, select the variable for which you want to see the structure defined in the model.
  2. Click the *View* button, the variable structure opens in a new window.

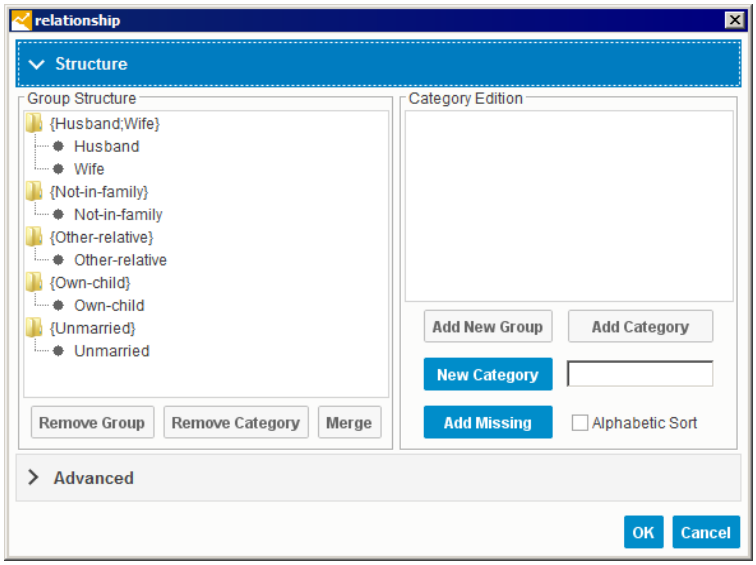


- If the variable has already been added to the list of variables located on the lower part of the panel:
  - Double-click the variable for which you want to see the structure defined in the model.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Create or Modify a Variable Structure

Double-click the Structure icon corresponding to the variable for which you want to edit or create the structure. The edit window opens. If the structure had been extracted from the variable statistics or from a model, the fields are already filled.



For details on how to use the Structure editor, see Structure by Type of Variables.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Structure for a Continuous Variable

The structure for a continuous variable is defined by several intervals each made of :

- a lower bound (*[ ]*) that can be either open or closed,
- a minimum value (*Minimum*),
- a maximum value (Maximum),
- a higher bound (*[ ]*) that can be either open or closed.

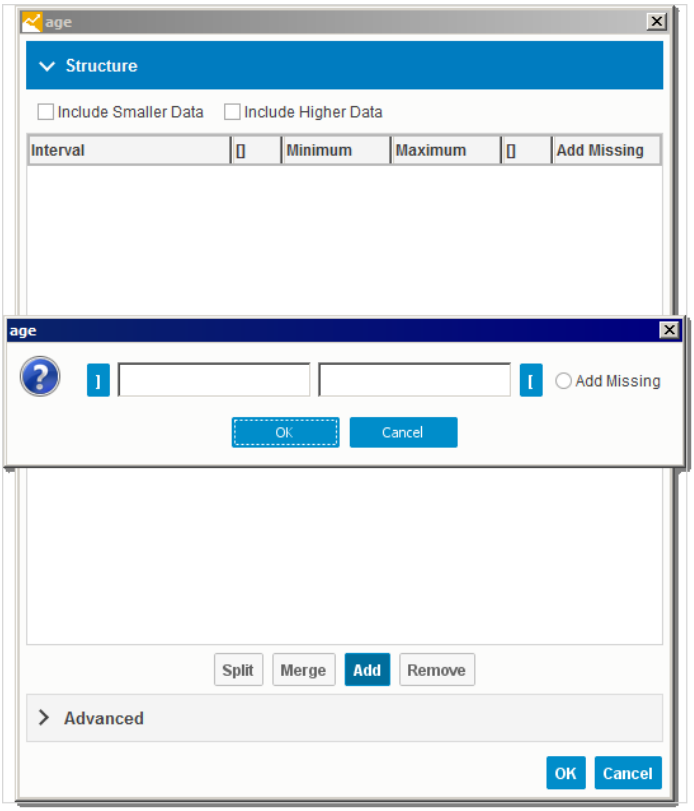
All intervals must be adjoining: there can be no gap or overlap between two intervals.

The option *Add Missing* allows you to indicate with which interval the missing values should be grouped.

The option *Include Smaller Data* allows you to include to the first interval any value smaller than its lower bound. In the same way, the option *Include Higher Data* allows you to include to the last interval any value higher than its higher bound.

☑ To Create a New Interval

1 Click the *Add* button to create a new interval. The edit window opens.



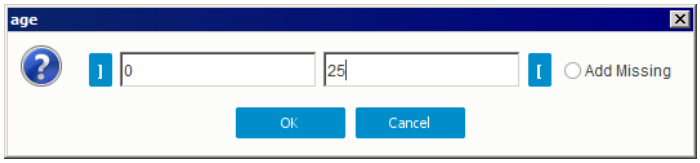
2 Select the lower bound type by clicking the *]* button.

3 Enter the minimum value for the interval in the left text field.

4 Enter the maximum value for the interval in the right text field.

5 Select the higher bound type by clicking the *[* button.

6 Check the option *Add Missing* if the missing values must be grouped with this interval.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

7 Click the *Yes* button to validate your interval.

☑ To Split an Interval

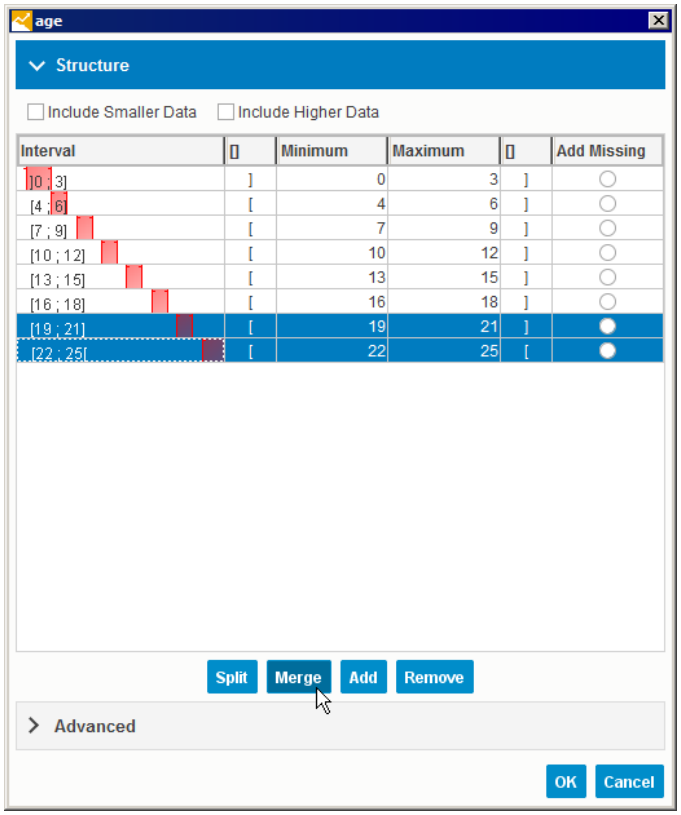
1 Select the interval to split.

2 Click the *Split* button. The selected interval is automatically split into two equal intervals.

☑ To Merge Two Intervals

1 Select the intervals to merge. You can only select adjoining intervals.

2 Click the *Merge* button.



☑ To Delete an Interval

1 Select one or more intervals. You can only select adjoining intervals.

2 Click the *Remove* button. The previous and next intervals are extended to include the values previously contained in the deleted intervals, so that no gap is left between intervals.

Structure for an Ordinal Variable

The structure for an ordinal variable is similar to that of a continuous variable with the exception of the bounds which are always closed and cannot be modified.

Caution

The structure for an ordinal string variable cannot be edited.

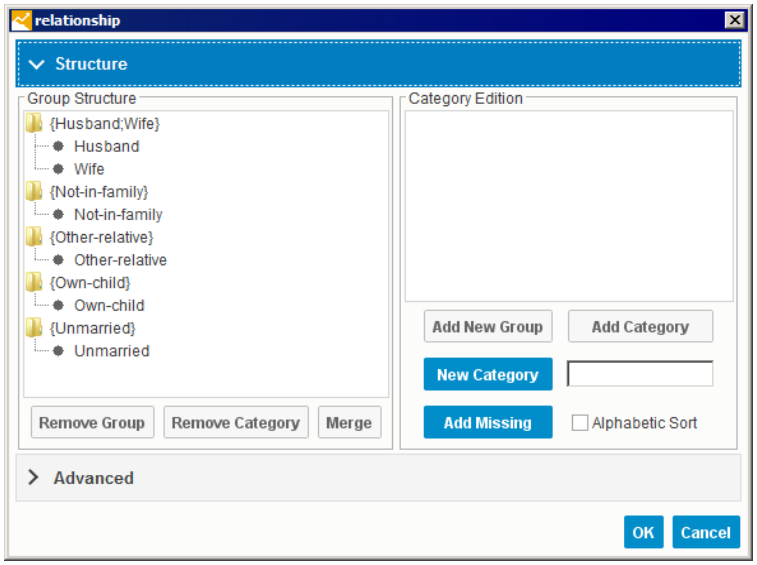
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Structure for a Textual Variable

The structure for a textual variable cannot be edited.

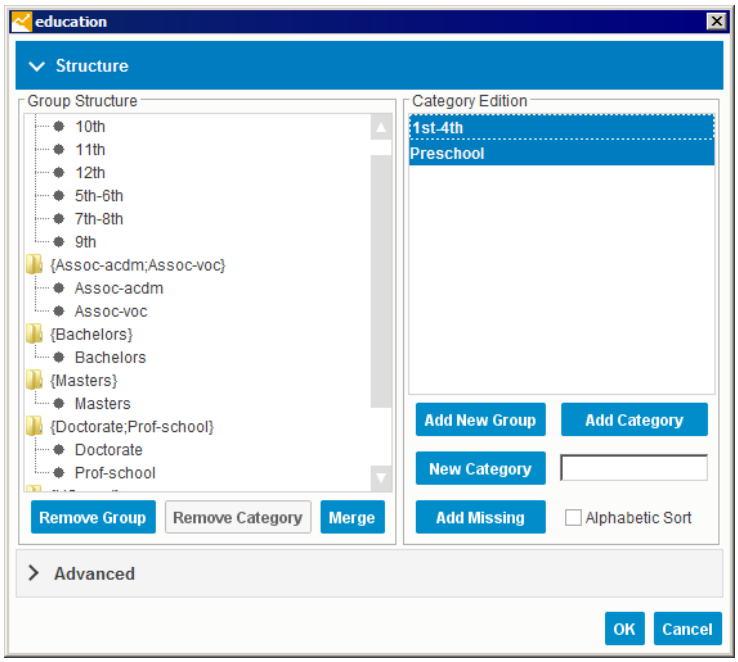
Structure for a Nominal Variable

The structure for a nominal variable is made of groups containing the variable categories.



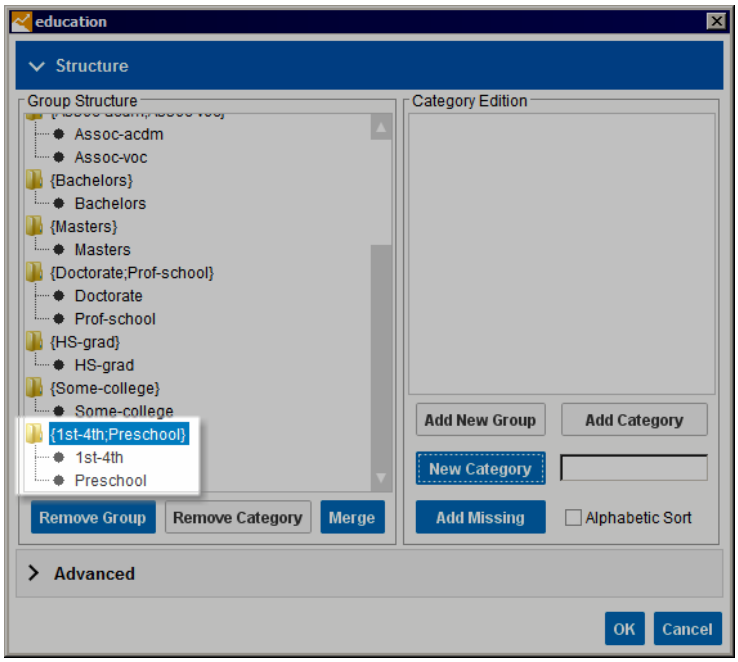
☑ To Create a New Category Group

1 In the list *Category Edition*, select the categories you want to add in a new group. Use the Ctrl key to select several categories.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

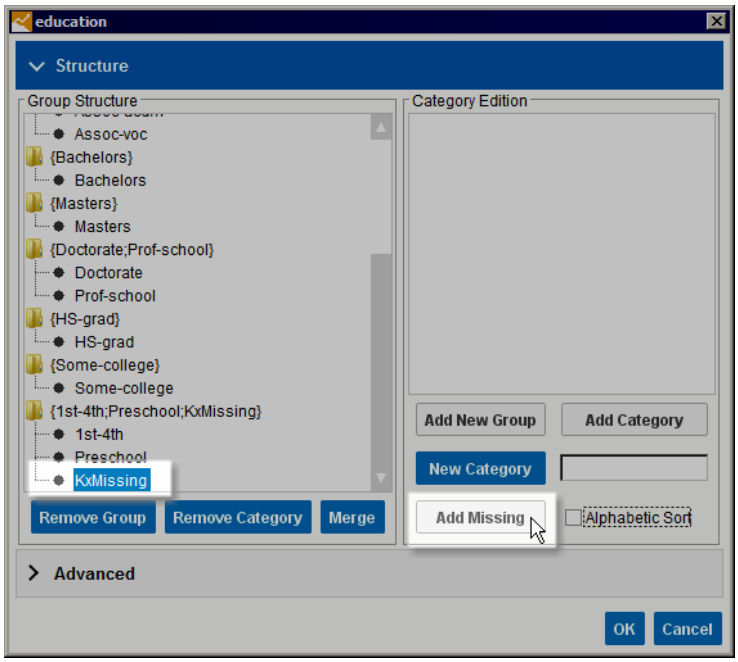
2 Click the button *Add New Group*. A group containing the selected categories is created in the list *Group Structure*.



☑ To Include Missing Values in a Group

1 In the list *Group Structure*, select the group in which you want to add the missing values.

2 Click the button *Add Missing* located under the list *Category Edition*. The `KxMissing` category, which represents the missing values, is added to the selected group and the button *Add Missing* is deactivated. As any category, the `KxMissing` category can only belong to one group at a time.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Create a New Category

1. In the field right of the button *New Category*, enter the name of the category to add.
2. Click the button *New Category*. The category is created in the list *Category Edition*.

☑ To Add Categories to a Group

1. In the list *Category Edition*, select the category (or categories) to add to a group.
2. In the list *Group Structure*, select the group in which you want to add the selected categories.
3. Click the button *Add Category*.

☑ To Delete a Group

1. In the list *Group Structure*, select the group to delete.
2. Click the button *Remove Group*. All the categories belonging to this group are re-added to the list *Category Edition*.

☑ To Remove a Category from a Group

1. In the list *Group Structure*, select the category or categories you want to remove from the group.
2. Click the button *Remove Category*. The selected categories are removed from the group and re-added to the list *Category Edition*.

Working Without any Defined Structure

If you let the structure as undefined, SAP InfiniteInsight® using consistent coder automatically determines the categories grouping depending on their interaction with the target variable. You can configure two parameters in this case:

- the band count for continuous variables
- InfiniteInsight® Modeler - Data Encoding optimal grouping for all variables.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Band Count for Continuous Variables

When you work with no defined structure, you can set the band count for continuous variables. The allowed values for this parameter are between 1 and 20.

The population is thus divided into as many segments of similar size. These segments are used to build descriptive statistics, particularly the distribution of target variables for each segment, which affects the coding of the variable with respect to target variables.

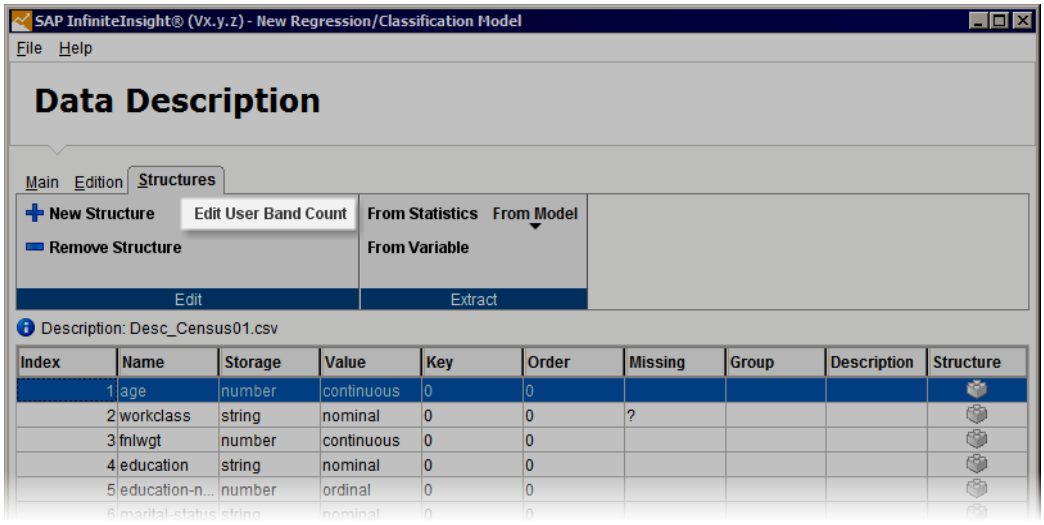
The band count has an influence on the calculation of the predictive power: the more there are segments, the more accurate is the calculation of the predictive power for the explanatory variable.. However, this influence is very small.

☑ To Edit the Band Count for Continuous Variables

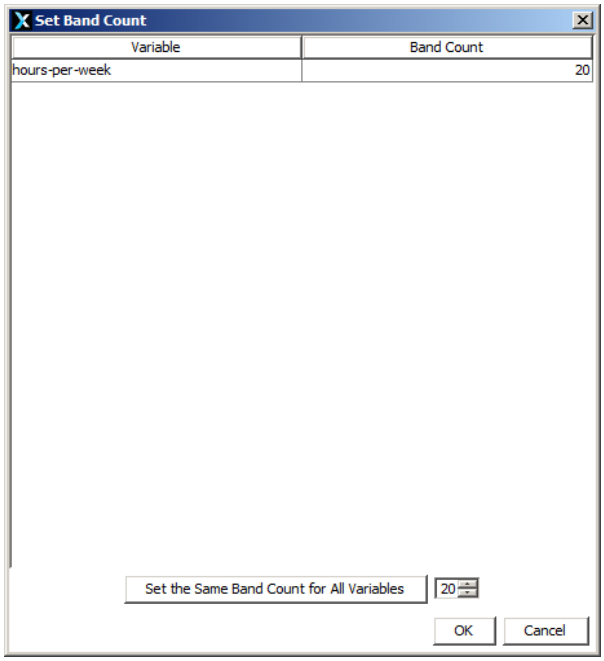
1 Click the row corresponding to the continuous variable to be edited.

2 Go to the *Structures* tab of the ribbon.

3 Click Edit User *Band Count*.



4 A screen is displayed as below:



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

If you want to...	Then...
modify the band count for all the continuous variables of the model	1 Type in the desired band count in the field at the bottom of the panel 2 Click *Set the Same Band Count for All Variables*. 3 Click *OK*.
modify the band count for the variable being edited	1 Type in the desired band count in the column *Band Count* at the top of the panel 2 Click *OK*.

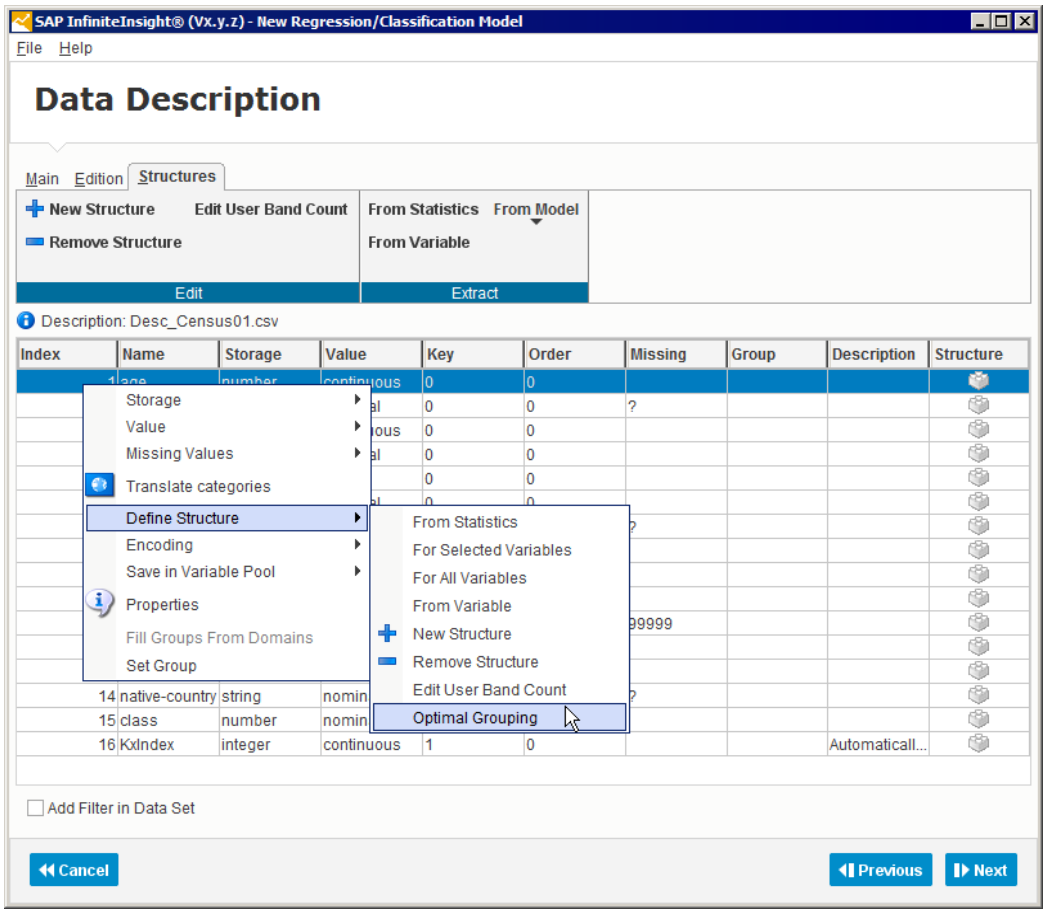
Optimal Grouping for All Variables

When working with a defined structure, if want to keep your categories as they are defined for the model building, you must disable this option.

If not or if you work with no defined structure, *Optimal Grouping* allows in a large number of cases to increase the prediction confidence of the model with a minimal loss of predictive power. Where possible, similar adjacent segments are gathered to reduce artifacts between the estimation and validation data sets.

☑ To Enable InfiniteInsight® Modeler - Data Encoding Optimal Grouping for All Variables

1. Right click the row corresponding to the variable to be edited.
2. Select *Define Structure*.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

3 Select *Optimal Grouping*, in such a way that the option is checked.

Filtering the Data Set

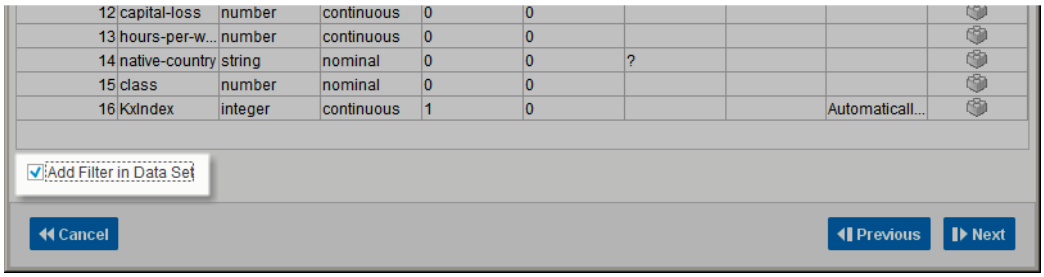
In order to accelerate the learn process and to optimize the resulting model, you can apply a filter to your data set.

For this scenario

Do not use the filtering option.

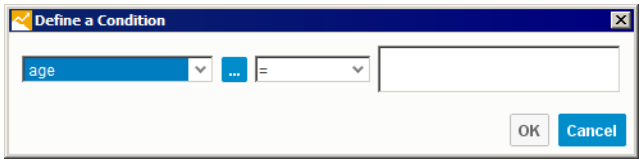
To Filter a Data Set

1 Check the option *Add a Filter in Data Set*.
2 Click *Next*.


12 capital-loss number continuous 0 0
13 hours-per-w... number continuous 0 0
14 native-country string nominal 0 0 ?
15 class number nominal 0 0
16 KxIndex integer continuous 1 0 Automaticall...
Add Filter in Data Set
Cancel
Previous
Next

To Add a Condition

1 Click the button *Add a Condition*.
The window *Define a Condition* opens.



2 Choose a variable in the first list.
3 Choose an operator in the second list.
4 Indicate a value in the third list:
- For a variable with number storage, type a value.
- For a variable with string storage, choose a variable in the list. If the list is empty click the button , to extract the variable categories.

5 Click *OK*.

> **Note**
>
> You can edit a condition by double-clicking it.

To Add a Logical Conjunction

Click the button *Add Logic 'And'* or the button *Add Logic 'Or'*.

> **Note**
>
> You can change a conjunction by double-clicking it.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Change the Order

You can change the order of the nodes to accelerate the filtering process by setting the conditions with the highest probability to be false at the top of the list.

1 Select the node you want to move up or down.

2 Use the buttons and to change its position in the filter.

☑ To Delete a Node

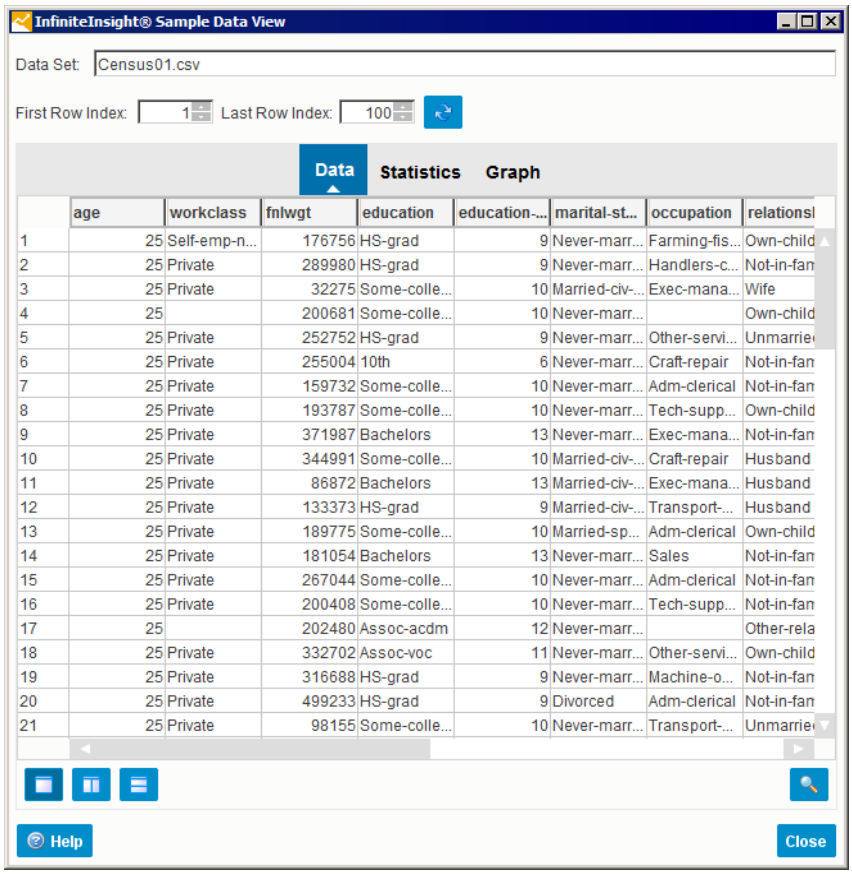
1 Select the node you want to delete.

2 Click the button *Remove Selected Node*.

☑ To Display the Filtered Data Set

You can visualize the data set that you will obtain after the application of the filter.

Click the button *View Data*.
A pop-up window opens.



	age	workclass	fnlwgt	education	education-...	marital-st...	occupation	relationsl
1	25	Self-emp-n...	176756	HS-grad	9	Never-marr...	Farming-fis...	Own-child
2	25	Private	289980	HS-grad	9	Never-marr...	Handlers-c...	Not-in-fan
3	25	Private	32275	Some-colle...	10	Married-civ-...	Exec-mana...	Wife
4	25		200681	Some-colle...	10	Never-marr...		Own-child
5	25	Private	252752	HS-grad	9	Never-marr...	Other-servi...	Unmarrie
6	25	Private	255004	10th	6	Never-marr...	Craft-repair	Not-in-fan
7	25	Private	159732	Some-colle...	10	Never-marr...	Adm-clerical	Not-in-fan
8	25	Private	193787	Some-colle...	10	Never-marr...	Tech-supp...	Own-child
9	25	Private	371987	Bachelors	13	Never-marr...	Exec-mana...	Not-in-fan
10	25	Private	344991	Some-colle...	10	Married-civ-...	Craft-repair	Husband
11	25	Private	86872	Bachelors	13	Married-civ-...	Exec-mana...	Husband
12	25	Private	133373	HS-grad	9	Married-civ-...	Transport-...	Husband
13	25	Private	189775	Some-colle...	10	Married-sp...	Adm-clerical	Own-child
14	25	Private	181054	Bachelors	13	Never-marr...	Sales	Not-in-fan
15	25	Private	267044	Some-colle...	10	Never-marr...	Adm-clerical	Not-in-fan
16	25	Private	200408	Some-colle...	10	Never-marr...	Tech-supp...	Not-in-fan
17	25		202480	Assoc-acdm	12	Never-marr...		Other-rela
18	25	Private	332702	Assoc-voc	11	Never-marr...	Other-servi...	Own-child
19	25	Private	316688	HS-grad	9	Never-marr...	Machine-o...	Not-in-fan
20	25	Private	499233	HS-grad	9	Divorced	Adm-clerical	Not-in-fan
21	25	Private	98155	Some-colle...	10	Never-marr...	Transport-...	Unmarrie

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Save a Filter

You can save the filter you created to be able to reuse it at a later moment without being obliged to recreate the same conditions.

1. Click the button *Save Filter*.
   A pop-up window is displayed.
2. In the list *Data Type*, select the format in which you want to save the filter.
3. Use the *Browse* button located on the right of the *Folder* field to select the folder or database where you want to save the filter.
4. In the *Description* field, enter the name of the file or table in which you want to save the filter.
5. Click *OK*.

☑ To Load an Existing Filter

To apply a filter to the data set, you can use a file created during a previous use of the data set in SAP InfiniteInsight®.

1. Click the button *Load Existing Filter*.
   A pop-up window is displayed.
2. Use the list *Data Type* to select the format of the filter.
3. Use the *Browse* button located on the right of the *Folder* field to select the folder or the database in which the filter is stored.
4. Use the *Browse* button located on the right of the *Description* field to select the file or the table containing the filter.
5. Click *OK*.

Translating the Variable Categories

You can translate the categories of a nominal variable, save the translation or load an existing translation. This translation has no influence on the variable structure, which has to be set according to the original values of the variable.

> **Note**
>
> The variable "Target Key", that is used in the advanced settings for example, does not take into account the translation when displaying the possible values of this variable.

☑ To Translate the Variable Categories

1. Click a nominal variable to translate its categories.
2. Go to the *Edition* tab of the ribbon and click the option *Translate Categories*. A new window appears.
3. Choose into which languages you want to translate. By default, the language of the user interface is displayed as a column.
4. Click the button to extract the variable categories from the data set.
5. Translate the categories.

> **Note**
>
> You do not need to fill all fields.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

6 Click *OK*.

☑ To Save the Categories Translation

1 Translate the variable categories as described above.
2 Click the *Save* button.
3 Choose a *Data Type*.
4 Select a *Folder*.
5 Enter a *Name* for the file or table.
6 Click *OK*.

☑ To Load an Existing Translation File

1 Click a nominal variable to translate its categories.
2 Go to the *Edition* tab of the ribbon and click the option *Translate Categories*. A new window appears.
3 Click the *Load* button.
4 Select the format of the translation in the list *Data Type*.
5 Use the *Browse* button located on the right of the *Folder* field to select the folder or the database in which the description is stored.
6 Use the *Browse* button located on the right of the field *Table or File* to select the file or the table containing the description.
7 Click *OK*.
8 Click the button *Update* to refresh the display of the categories.
9 If the list of columns is not named correctly, use the *Advanced Settings* (see next paragraph) to set a header line and update again.
10 Map the language names with those from the loaded translation, by clicking the categories and choosing the corresponding language in the contextual menu.
11 Click *OK*.

Selecting Variables

Once the training data set and its description have been entered, you must select the following variables:

- one or more Target Variables (on page 31)
- possibly a Weight Variable (on page 34)
- the Explanatory Variables (on page 32)

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

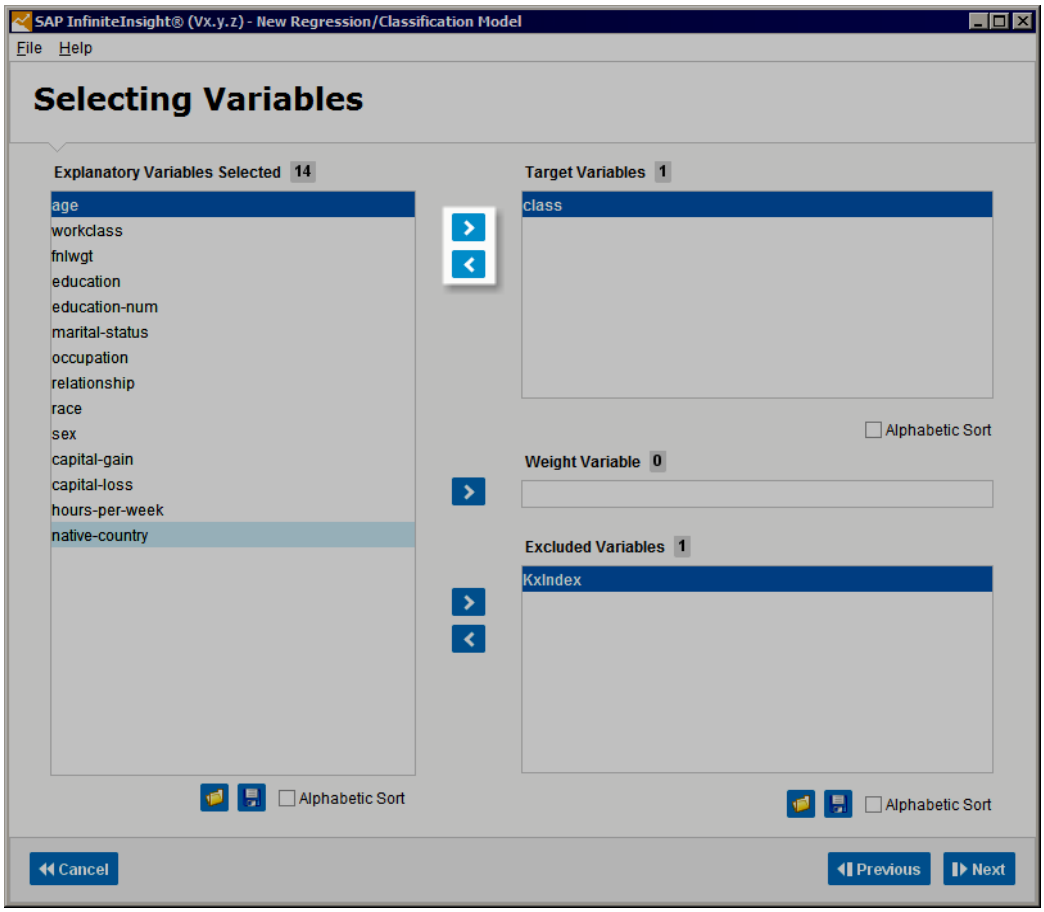
Target Variables

For this Scenario

Select the variable *Class* as your **target variable**, that is, the variable that indicates the probability of an individual responding in a positive or negative manner to your campaign.

To Select Targets Variables

1 On the screen *Selecting Variables*, in the section *Explanatory variables selected* (left hand side), select the variables you want to use as Target Variables.



> **Note**
>
> On the screen *Selecting Variables*, variables are presented in the same order as that in which they appear in the table of data. To sort them alphabetically, select the option *Alphabetic sort*, presented beneath each of the variables list.

2 Click the button > located on the left of the screen section *Target Variables* (upper right hand side).
The variable moves to the screen section *Target Variables*.
Also, select a variable in the screen section *Target Variables* and click the button < to move the variables back to the screen section *Explanatory variables selected*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Weight Variable

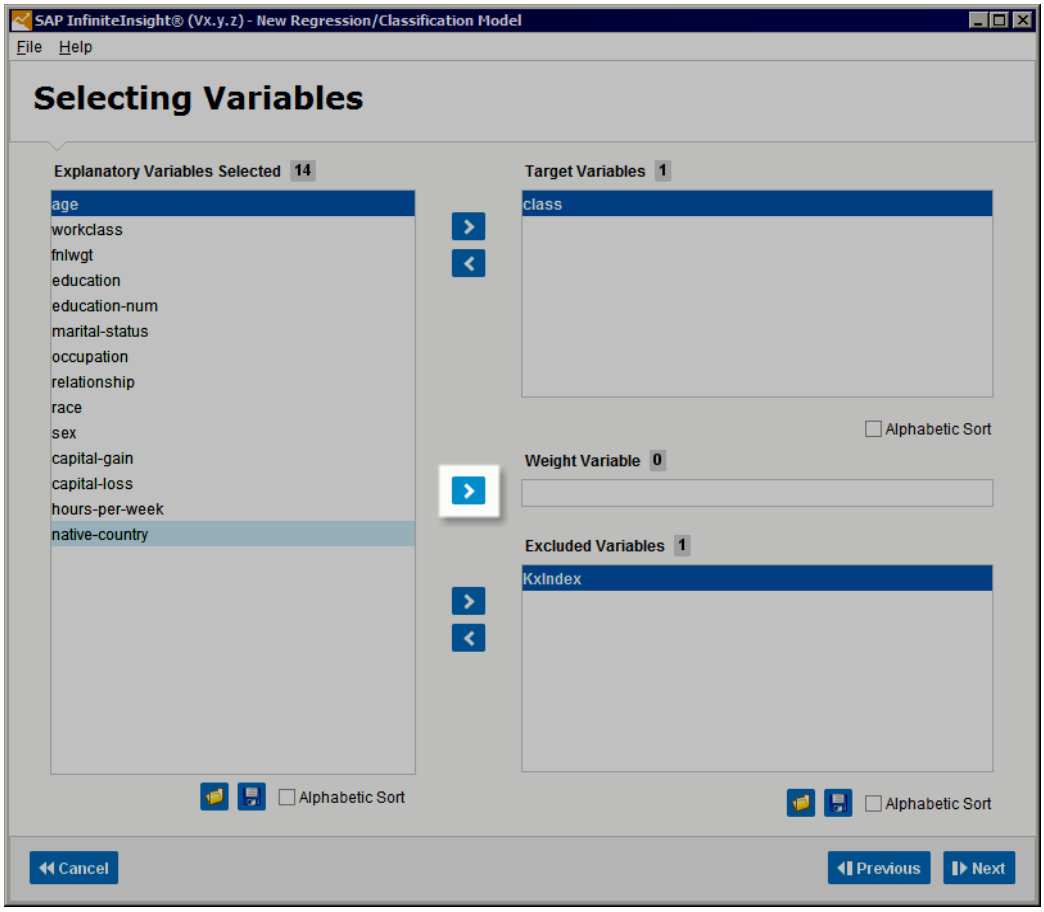
Selecting a **Weight Variable** enables to set the **Weight Quantum** option available in the **Advanced Model Parameters**.

For this Scenario

Do **not** select a weight variable.

To Select a Weight Variable

1 On the screen *Selecting Variables*, in the section *Explanatory variables selected* (left hand side), select the variables you want to use as a Weight Variable.



> **Note**
>
> On the screen *Selecting Variables*, variables are presented in the same order as that in which they appear in the table of data. To sort them alphabetically, select the option *Alphabetic sort*, presented beneath each of the variables list.

2 Click the button > located on the left of the screen section *Weight Variable* (middle right hand side).
The variable moves to the screen section *Weight Variable*.
Also, select a variable in the screen section *Weight Variable* and click the button < to move the variables back to the screen section *Explanatory variables selected*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Explanatory Variables

By default, and with the exception of key variables (such as *KxIndex*), all variables contained in your data set are taken into consideration for generation of the model. You may exclude some of these variables.

For the first analysis of your data set, we recommend that you **retain all variables**. It is particularly important to retain even the variables that seem to have no impact on the target variable. If indeed these variables have no impact on the target variable, the model will confirm this. In the contrary case, the model will allow you to recognize previously unidentified correlations between these variables and the target variable. By excluding variables from the analysis based on simple intuition, you take the risk of depriving yourself of one of the greatest value-added features of SAP InfiniteInsight® models: the discovery of **non-intuitive information**.

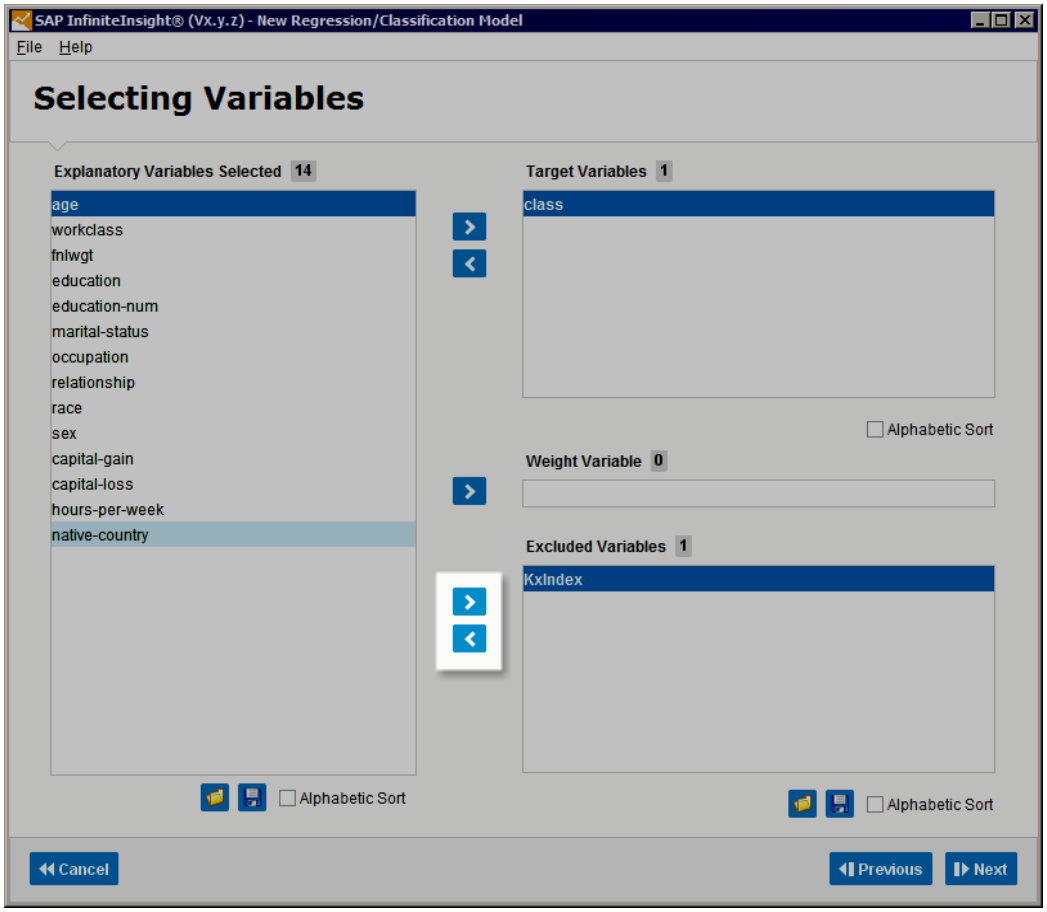
Depending on the results obtained from the first analysis, which included all of the variables of the data set, you can generate a second model by excluding the variables too closely correlated with the target variable.

For this Scenario

- Exclude the variable *KxIndex*, as this is a key variable. Since the initial data set does not contain a key variable, SAP InfiniteInsight® feature generated KxIndex automatically.
- Retain all the other variables.

To Exclude Variables from Data Analysis

1 On the screen *Selecting Variables*, in the section *Explanatory Variables Selected* (left hand side), select the variable to be excluded.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

> **Note**
>
> On the screen *Selecting Variables*, variables are presented in the same order as that in which they appear in the table of data. To sort them alphabetically, select the option *Alphabetic sort*, presented beneath each of the variables list.

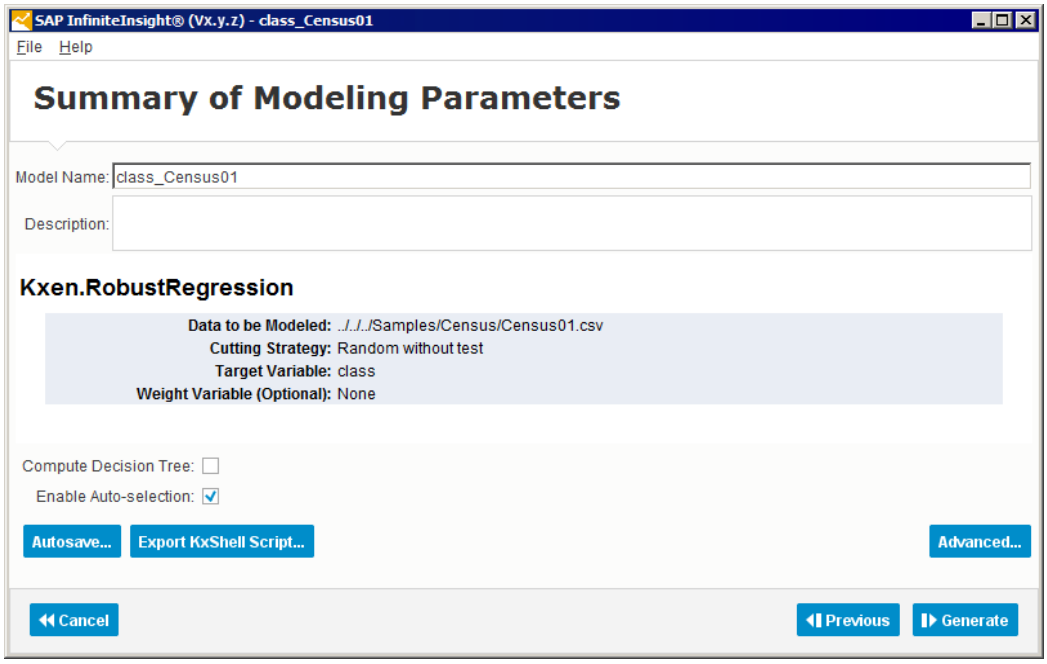
2 Click the button > located on the left of the screen section *Excluded Variables* (lower right hand side).
The variable moves to the screen section *Excluded Variables*.
Also, select a variable in the screen section *Excluded Variables* and click the button < to move the variables back to the screen section *Explanatory Variables Selected*.

> **Note**
>
> By default, any variable defined as a key is put in the *Excluded Variables*. However, the user has the possibility to move a key variable in the *Explanatory Variables Selected* if he wants this variable to have this role.

3 Click *Next*.
The screen *Parameters of the Model* appears.

Checking Modeling Parameters

The screen *Summary of Modeling Parameters* allows you to check the modeling parameters just before generating the model.



> **Note**
>
> The screen *Summary of Modeling Parameters* contains an *Advanced* button. By clicking this button, you access the screen *Specific Parameters of the Model*. For more information about these parameters,Setting the Advanced Parameters (on page 99).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

- The *Model Name* is filled automatically. It corresponds to the name of the target variable (`class` for this scenario), followed by the underscore sign ("_") and the name of the data source, minus its file extension (`Census01` for this scenario).
- You have the possibility to **display the results** generated by K2R as **a decision tree** based on the five most contributive variables. To activate this option, check the box *Compute Decision Tree*.
- The *Autosave* button allows you to activate the feature that will automatically save the model once it has been generated. When the autosave option is activated, a green check mark is displayed on the *Autosave* button.

Activating the Autosave Option

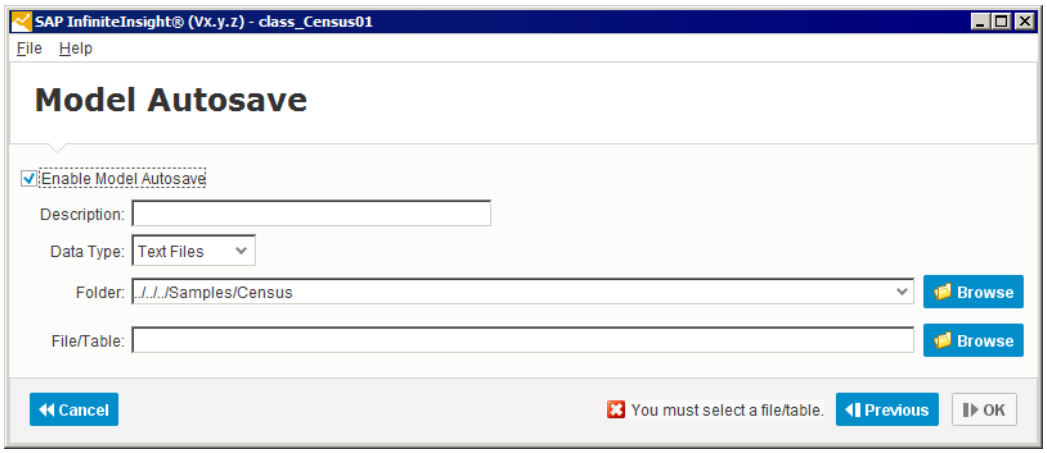
The panel *Model Autosave* allows you to activate the option that will automatically save the model at the end of the generation process and to set the parameters needed when saving the model.

☑ To Activate the Autosave Option

1 In the panel *Summary of Modeling Parameters*, click the *Autosave* button.

The panel *Model Autosave* is displayed.

2 Check the option *Enable Model Autosave*.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

3 Set the parameters listed in the following table.

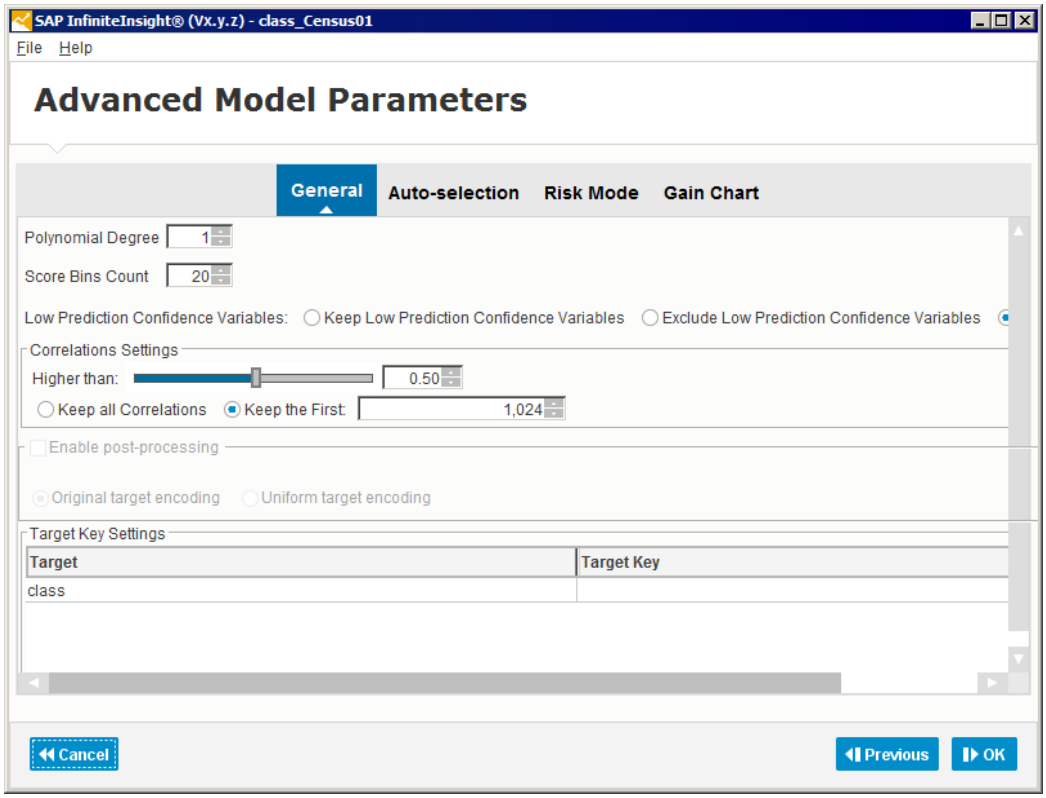
Parameter	Description
Model Name	This field allows you to associate a name with the model. This name will then appear in the list of models to be offered when you open an existing model.
Description	This field allows you to enter the information you want, such as the name of the training data set used, the polynomial degree or the performance indicators obtained. This information could be useful to you later for identifying your model. Note that this description will be used instead of the one entered in the panel *Summary of Modeling Parameters*.
Data Type	this list allows you to select the type of storage in which you want to save your model. The following options are available: ▪ *Text files*, to save the model in a text file. ▪ *Database*, to save the model in a database. ▪ *Flat Memory*, to save the model in the active memory. ▪ *SAS Files*, to save the model in a SAS compatible file for a specified version of SAS and a specified platform (SAS v6 or 7/8 for Windows or UNIX). ▪ *SAS Transport*, to save the model in a generic SAS compatible file.
Folder	Depending upon which option you selected, this field allows you to specify the ODBC source, the memory store or the folder in which you want to save the model.
File/Table	This field allows you to enter the name of the file or table that is to contain the model. The name of the file must contain one of the following format extensions `.txt` (text file in which the data is separated by tabs) or `.csv` (text file in which the data is separated by commas).

4 Click *OK*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Setting the Advanced Parameters

On the screen *Summary of Modeling Parameters*, click the *Advanced* button.
The screen *Advanced Model Parameters* appears.



General Tab

The *General* tab allows you to define the general settings of the model, that is, the degree of the model, the score bin count, the number of correlations to display and the target key value.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Defining the Degree of the Model (optional)

The model generated by InfiniteInsight® Modeler - Regression/Classification is represented by a polynomial. This polynomial may be of degree 1, 2, 3 or greater. By defining the polynomial degree, you will define the degree of complexity of the model.

It is greatly recommended that you always use a degree of "1" (default value) for the first analysis of a data set. Using a higher degree of polynomial does not guarantee that you will in all cases obtain a more powerful model.

For more information about the polynomial degree, Representation of a Model (see page 36).

For this Scenario

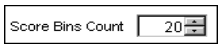
Keep the polynomial degree set to the default value – that is "1".

To Define the Degree of the Model

In the *Polynomial degree* field, enter the value corresponding to the degree of complexity of the model that you want to obtain.

Setting the Score Bins Count

This option allows you to define the numbers of bins to create for the score. This value must be set between 20 and 100 since a lower or higher number of bins would lead to poor model quality.



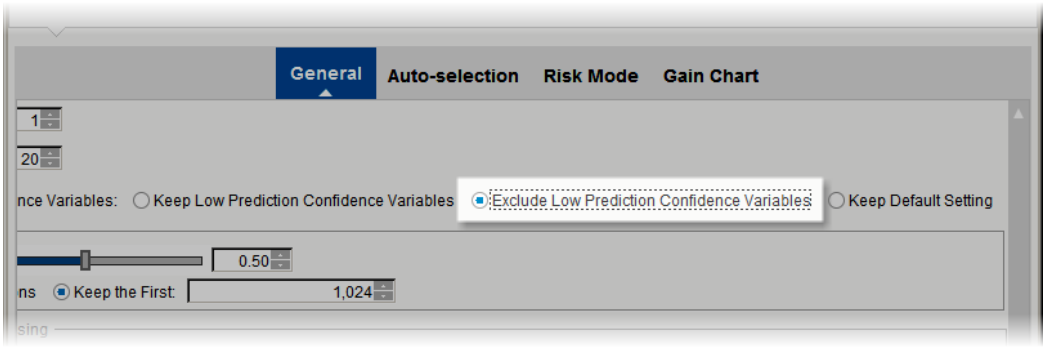
Exclusion of Low KR Variables

This option allows you to enable the exclusion of variables based on the value of their prediction confidence (KR). InfiniteInsight® uses an internally computed threshold to decide whether a variable has a low prediction confidence. This threshold depends mainly on the data set size and target distribution.

In versions prior to 6.1.0, InfiniteInsight® automatically excluded variables with low prediction confidence; since version 6.1.0 this behavior has been disabled by default. If you do not enable this feature, no variable will be excluded based on the value of its prediction confidence.

To Automatically Exclude Variables with Low Prediction Confidence

Check the option *Exclusion of Low Prediction Confidence Variables*.

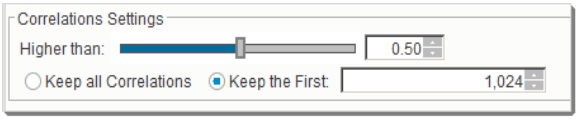


© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Defining the Number of Correlations to Display

The section *Correlations Settings* allows you to set the parameters for the *Correlation* debriefing panel. That is, to select how many correlations should be displayed in that panel.

To say that variables are correlated implies that they each contribute some of the same information with respect to the target variable. A correlation contains two variables and a correlation rate. When you modify the number of correlations to display, the engine excluded the ones with the lowest correlation rate, thus keeping only the more significant ones.

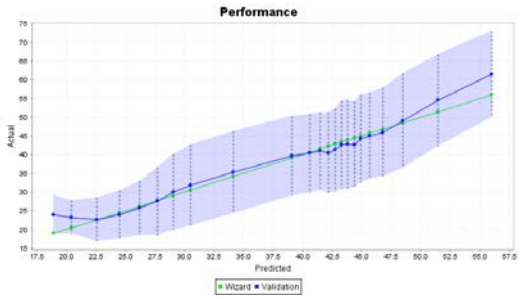
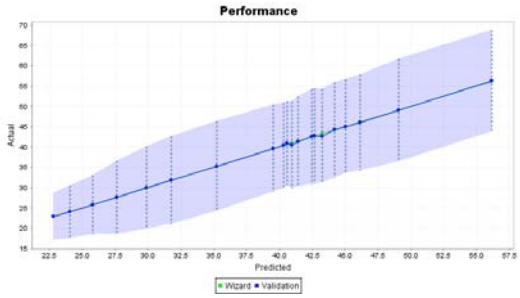
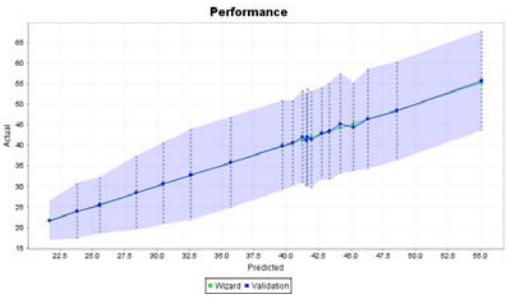


© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Enabling the Post-Processing

This section allows setting some regression parameters according to three strategies. This option can only be activated when the model contains at least one continuous target variable.

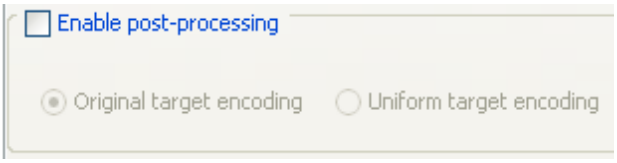
The description of these strategies and an example of performance curve for each strategy are provided in the table below.

Regression Strategy	Description	Example of Performance Curve
Without post-processing	The first strategy consists in disabling the regression post-processing during the learning model phase in order to create a regression similar to the one used in versions prior to 3.3.2. In this case, a standard regression is performed. No special improvement is made to the final scores. Original target values are used and raw score values are produced as outputs.	
With original target encoding	The second strategy, which applies to regressions using a post-processing, consists in using the original target value during the learning model phase to compute regression coefficients. The result of the regression is then transformed to align target segment means and score segment means in the post-processing phase. **Note** This is the default strategy used in InfiniteInsight™.	
With uniform target encoding	The last strategy, which applies to regressions using a post-processing, consists in using first an encoded target value instead of the original target value during the learning model phase in order to have a uniform distribution: it is the pre-processing phase. Then, regression coefficients are computed and scores are transformed in the original target space during the post-processing phase. **Note** This strategy is to be preferred when the default strategy does not produce models with enough quality, which is often the case with very skewed target distributions.	

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ Regression Without Post-processing

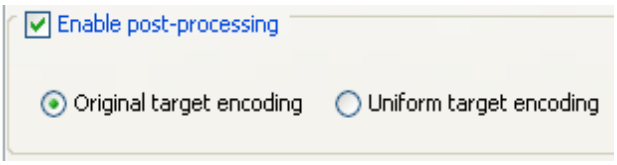
Uncheck the option *Enable Post-processing*.



> **Note**
>
> It is not possible to change the target encoding strategy when the post-processing is disabled.

☑ Regression with Original Target Values

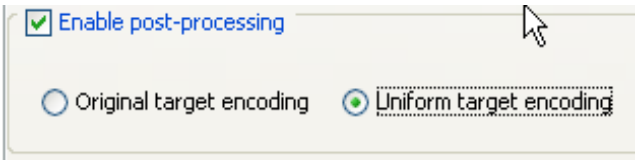
1. Check the option *Enable Post-processing*.
2. Select the radio button *Original target encoding*.



> **Note**
>
> This regression strategy corresponds to regressions used in SAP InfiniteInsight® from version 3.3.1 to version 3.3.6. This strategy is set by default.

☑ Regression with Uniform Target Encoding

1. Check the option *Enable Post-processing*.
2. Select the radio button *Uniform target encoding*.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Defining the Target Key Values

For the binary targets you have the option to select which value is the key category for each target. By default, the category selected by SAP InfiniteInsight® is the least represented in the data set.

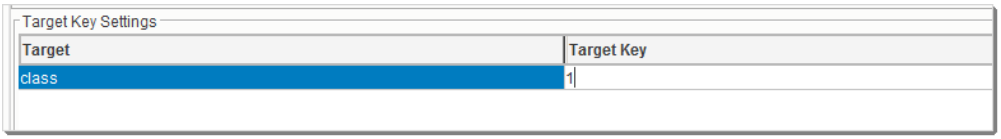
The *Advanced Model Parameters* screen lists all the binary targets of the current model allowing you to define the key category for each target, that is the expected value of the target.

In this scenario

Do not define a value for the target variable. SAP InfiniteInsight® will automatically select "1" as the key category for the Class variable.

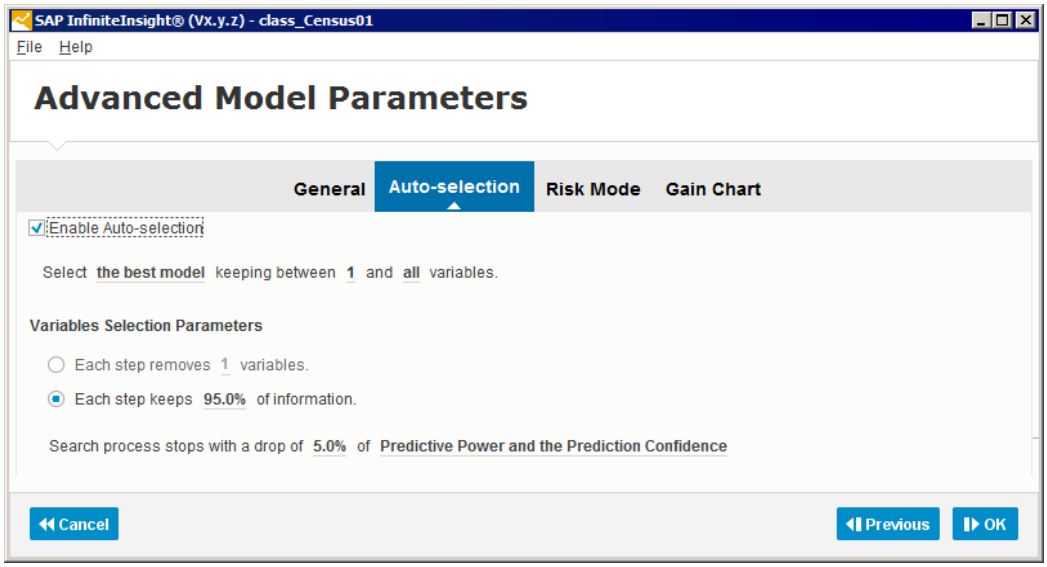
To Define the Key Category Value for a Target Variable

In the *Target Key* field corresponding to the chosen target, enter the key value.



Auto-selection Tab

The *Auto-selection* tab allows you to define the parameters of the automatic variable selection.



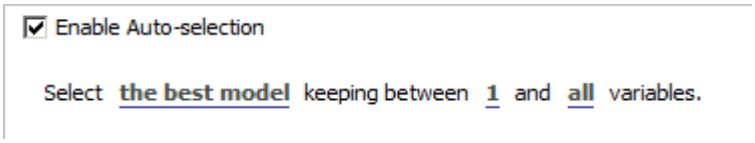
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Setting the Variables Auto-Selection

The section *Auto-selection* allows you to automatically reduce the number of variables in the model in relation to quality criteria. This selection is done by successive iterations. There are two selection modes, one based on the number of variables to keep, and the other on the amount of information that should be kept. In this instance, the information is the sum of the variables contributions.

☑ To Use the Auto-selection

Check the box *Enable Auto-Selection*. The corresponding options are activated.



By default the parameters are set to: "*Select the best model keeping between 1 and all variables*".

Any parameter that can be changed is marked as a hyperlink (blue, underlined).

Choosing the Selection Mode

☑ To Select the Selection Mode

1. Click the link indicating the type of model to keep. For example, *the best model* in the sentence "*Select the best model keeping between 1 and all variables.*"
   A drop-down menu is displayed offering the following options:
   - *the best model*
   - *the last model*
2. Select the desired option.
3. Click *OK*.

☑ To Select the Number of Variables

This criterion is mandatory and allows fixing the minimum and the maximum number of variables in the final model.

1. In the sentence defining the number of variables "*Select the best model keeping between 1 and all variables*", select the minimum number of variables (for example, *1 variable*) and the maximum number of variables (for example, *all variables*) .
2. For the minimum number of variables, a slide is displayed ranging from 1 to the total number of variables in the model.

   Move the cursor on the slide to select the quantity of your choice.

   For the maximum number of variables you can either confirm the minimum number selected previously by choosing "Keep all variables" or choose a maximum number of variables.
3. Click *OK*.

Choosing the Stopping Criteria

You can choose between two variables selection parameters:

- *Each step removes 1 variable.*

This option allows you to set the number of variables that should be excluded at each iteration.

- *Each step keeps 95.0% of information.*

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

This option allows you to set the amount of information that should be kept at each iteration (thus limiting the loss of information).

Select the desired option.

☑ To Select the Number of Variables

1. Click the link indicating the number of variables in the sentence "*Each step removes 1 variable*." A slide is displayed ranging from 1 to the total number of variables in the model.
2. Move the cursor on the slide to select the number of your choice.
3. Click *OK*.

☑ To Select the Information Amount

1. Click the link indicating the amount of information to keep in the sentence "*Each step keeps 95.0% of information*." A slide is displayed.
2. Move the cursor on the slide to select the quantity of your choice.
3. Click *OK*.

☑ To Set the Authorized Quality Loss

The quality loss can be set in the sentence "*Search process stops with a drop of 1.0% of KI and KR*".

1. Click the link indicating the percentage of loss (for example, *5.0%*). A slide is displayed.
2. Select the maximum percentage of authorized quality loss with the cursor.
3. Click *OK*.
4. Click the quality criterion. A drop-down list is displayed offering the following options:
   - *Based on KI + 2 KR,* the quality loss is based on both the predictive power (KI) and twice the prediction confidence (KR).
   - *KI and the KR*, the quality loss is limited for both the predictive power (KI) and the prediction confidence (KR). It is the default value.
   - *KI*, the quality loss is limited for the predictive power (KI) only.
   - *KR*, the quality loss is limited for the prediction confidence (KR) only.
5. Select the option of your choice.
6. Click *OK*.

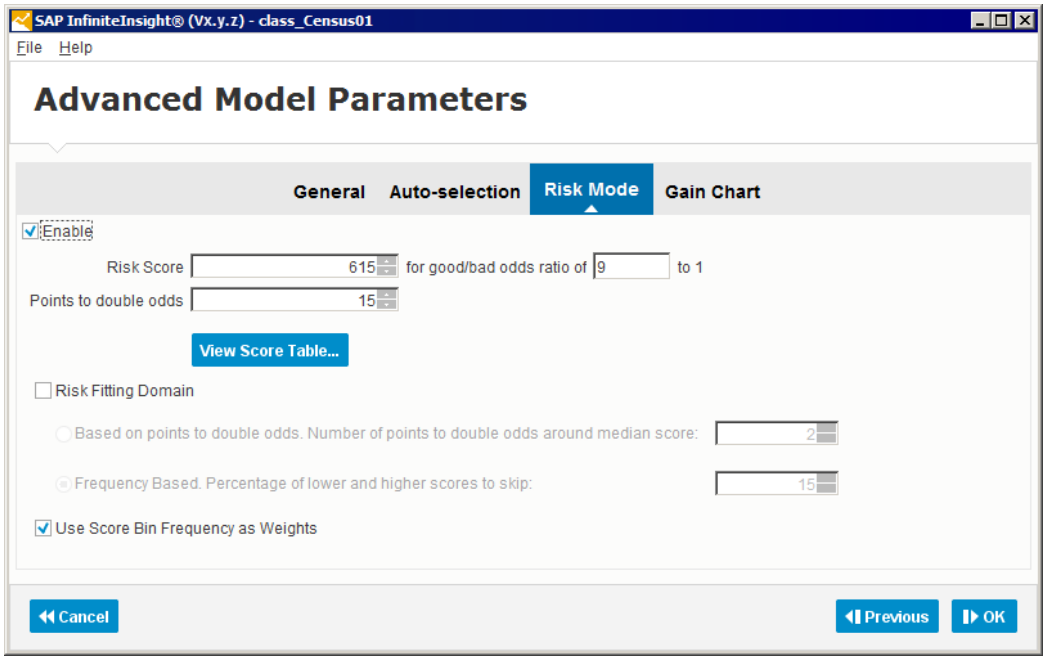
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Risk Mode Tab

This tab allows you to select a specific learning mode for your model.

☑ To Enable the Risk Mode

1 Select the tab *Risk Mode*.



2 Check the box *Enable*. The tab activates and the *Risk Mode* settings are displayed.

Setting Risk Mode

InfiniteInsight® **Risk Mode** allows advanced users to ask a classification model to translate its internal equation obtained with no constraints into a specified range of scores associated with good/bad odds ratio.

When this mode is activated, the different encodings that are used internally for continuous and ordinal variables are merged in a single representation, allowing a simpler view of the model internal equations. This is particularly useful when the usage of predictive model is subject to legal restrictions: the model equations are now simple enough to be understood by legal departments, and can be exposed, not only in programming language, as it was already the case before, but even in simple words.

The underlying technology is also used to display so called 'score cards'.

To use this mode, you need to choose:

- a *Risk Score* associated with a *Good/Bad Odds* ratio

  Note

  The odds ratio is the ratio between 'good' and 'bad', i.e. (1-p)/p where p is the probability of risk.

- the number of *Points to double the odds*

  Note

  The points to double the odd (PDO) are the number of risk points needed to double the odds ratio.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Example

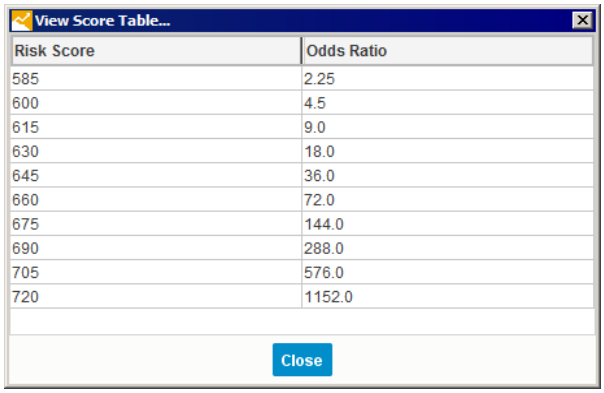
Considering a **Risk Score** equal to `615`, an odds ratio of `9 :1` and 15 points to double the odds. In this case, InfiniteInsight® will automatically re-scale the internal scores to scores in **Risk Mode** space and associates an odds ratio to each score in **Risk Mode** space.

In this Scenario

Do not activate the Risk mode.

To Define the Risk Mode Parameters

1. In the field *Risk Score*, enter the score you want to associate with a good/bad odds ratio.
2. In the field *for good/bad odds ratio of*, enter the ratio.
3. Indicate the increase of score points needed to double the odds in the field *Points to double odds*.
4. Click the button *View Score Table* to display the table of scores associated with the corresponding good/bad odds ratio.



Risk Fitting Domain

This option allows the user to control the way risk score fitting is performed, that is, how InfiniteInsight® fits its own scores to the risk scores.

The risk fitting has two modes:

- *PDO based*: the area equals `[Median Score - N*PDO ; Median Score + N*PDO]`. `N` (number of PDOs around the median score) must be specified by the user. By default, `N` is set to `2`.

  #### Note

  PDO stands for Points to double the odds.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

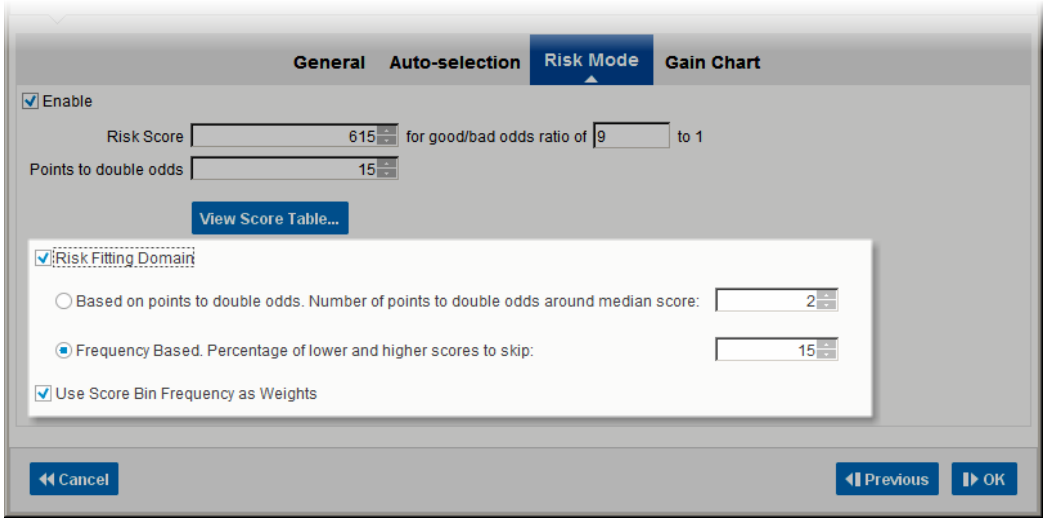
- *Frequency based*: the area equals `[Quantile(Freq) ; Quantile(1.0 - Freq)]`. The frequency of higher and lower scores to be skipped must be specified by the user. By default, the frequency is set to 15%.

If you do not check the box *Risk Fitting Domain*, the mode **Frequency based** will be used by default.

The fitting can be weighted or not.

☑ To Set the Risk Fitting Parameters

1 Check the box *Risk Fitting Domain*.



2 Select the mode you want to use.

3 Depending on the selected mode, set the appropriate value in the corresponding field.

4 If you want to use weights for the fitting, check the box *Use Score Bin Frequency as Weights*.

5.2.2 Step 2 - Generating and Validating the Model

Once the modeling parameters are defined, you can generate the model. Then you must validate its performance using the predictive power (KI) and the prediction confidence (KR):

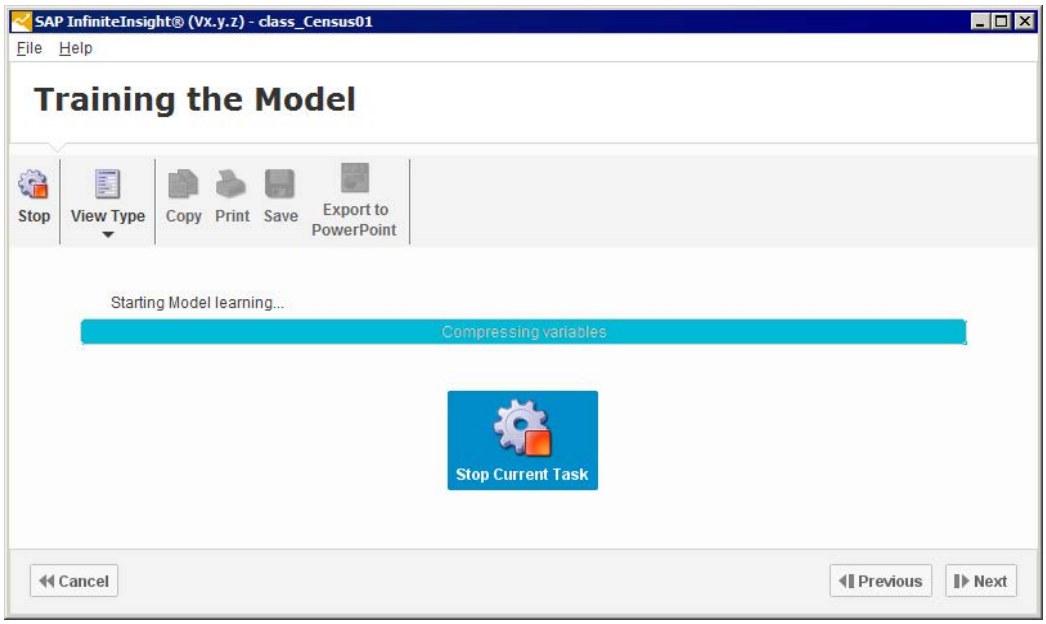
- If the model is sufficiently powerful, you can analyze the responses that it provides in relation to your business issue (see Step 3 - Analyzing and Understanding the Model Generated (see page 113)), and then apply it to new data sets (see Step 4 - Using the Model (see page 154)).
- Otherwise, you can modify the modeling parameters in such a way that they are better suited to your data set and your business issue, and then generate new, more powerful models.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

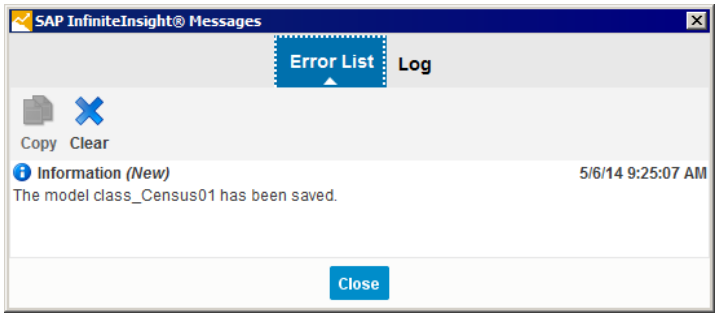
Generating the Model

☑ To Generate the Model

1 On the screen *Specific Parameters of the Model*, click the Generate button.
The screen *Training the Model* appears. The model is being generated. A progress bar allows you to follow the process.



2 If the *Autosave* option has been activated in the panel *Summary of Modeling Parameters*, a message is displayed at the end of the learning process confirming that the model has been saved.



3 Click *Close*.

4 Once the model has been generated, click *Next* to go to panel **Using the Model**.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Following the Progress of the Generation Process

There is two ways for you to follow the progress of the generation process:

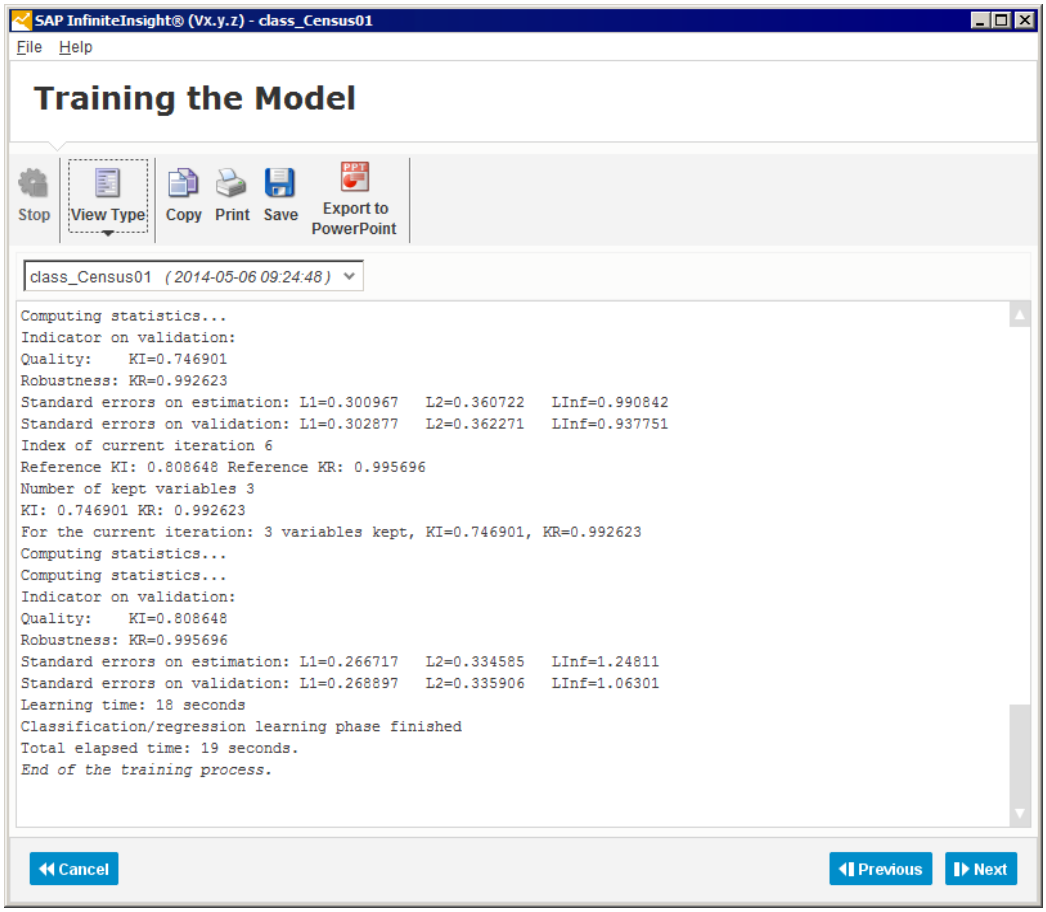
- The **Progress Bar** displays the progression for each step of the process. It is the screen displayed by default.
- The **Detailed Log** displays the details of each step of the process.

☑ To display the Progression Bar

Click *View Type* and select the (*Progression*) button.
The progression bar screen appears.

☑ To Display the Detailed Log

Click *View Type* and select the (*Log*) button.
The following screen appears.



☑ To Stop the Learning Process

1 Click the (*Stop Current Task*) button.
2 Click the *Previous* button.
The screen *Summary of Modeling Parameters* appears.
3 Go back to the section **Check Modeling Parameters**.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Validating the Model

Once the model has been generated, you must verify its validity by examining the performance indicators:

- The **predictive power** allows you to evaluate the **explanatory power** of the model, that is, its capacity to explain the target variable when applied to the training data set. A perfect model would possess a predictive power equal to 1 and a completely random model would possess a predictive power equal to 0.
- The **prediction confidence** defines the degree of robustness of the model, that is, its capacity to achieve the same explanatory power when applied to a new data set. In other words, the **degree of robustness** corresponds to the predictive power of the model applied to an application data set.

For more details on the model results, see section Model Summary.

For this Scenario

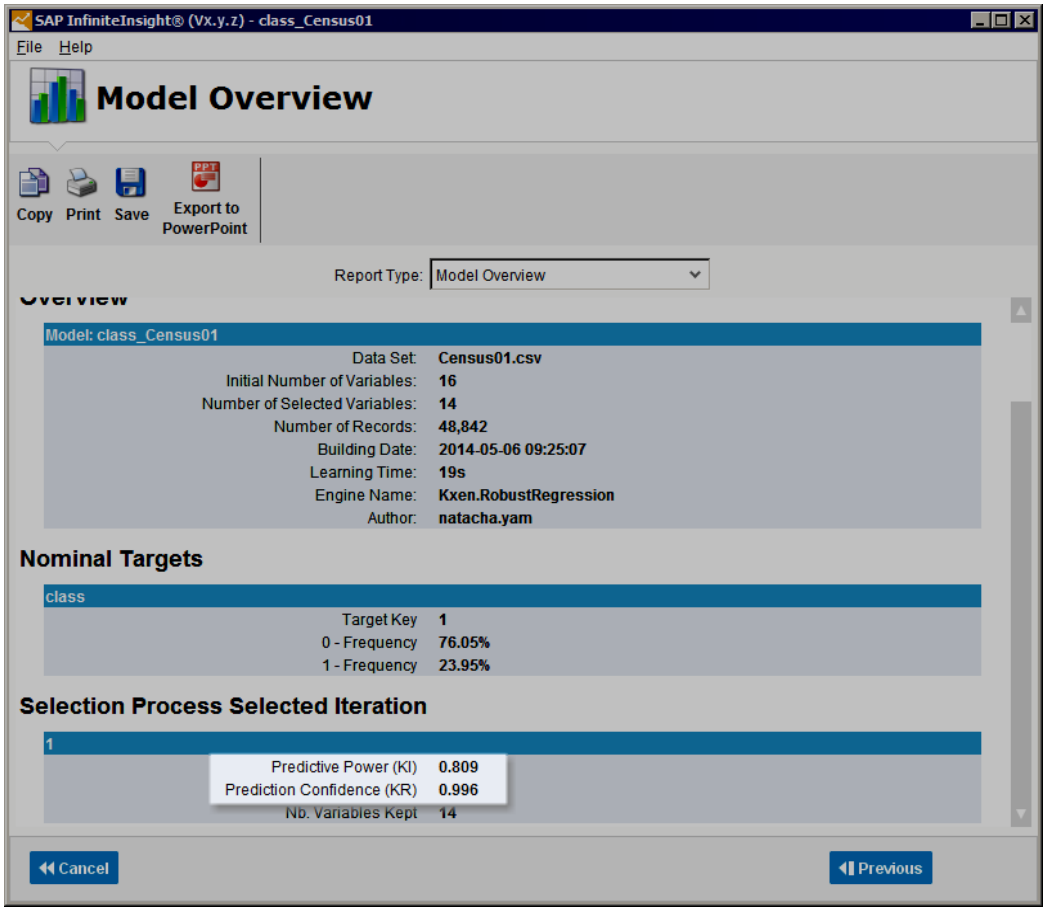
The model generated possesses:

- A quality indicator KI equal to 0.808,
- A robustness indicator KR equal to 0.992.

The model performs sufficiently well. You do not need to generate another.

To Validate the Model Generated

1 Verify the *Predictive Power (KI)* and *Prediction Confidence (KR)* of the model. These indicators are marked in red on the following figures.



a) If the performance of the model meets your requirements, go to **Step 3 "Analyzing and Understanding the Model Generated (see page 113)"**
b) Otherwise, go to the procedure **To Generate a New Model**.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2 You can also check other indicators provided in addition to *KI* and *KR* during the model generation. For example, you could view the *total elapsed time* required to generate the model and information on the *standard error* rate.

☑ To Generate a New Model

You have two options. On the screen *Training the Model*, you can:

- Either click the *Previous* button to return to the modeling parameters defined initially.
  Then you can modify the parameters one by one.
- Or click the *Cancel* button to return to the main screen of InfiniteInsight® modeling assistant.
  Then you must redefine all the modeling parameters.

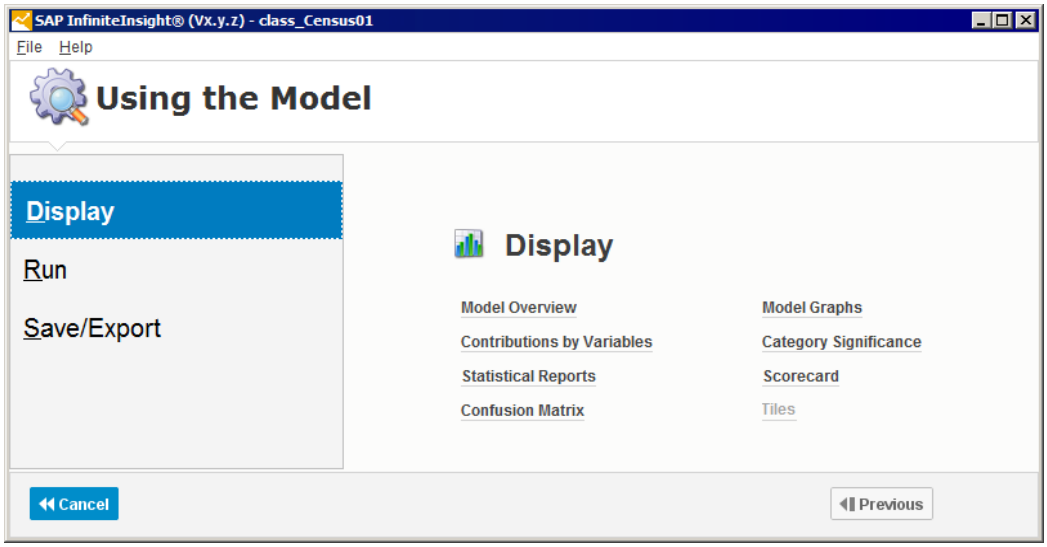
5.2.3 Step 3 - Analyzing and Understanding the Model Generated

The suite of plotting tools within SAP InfiniteInsight® allows you to analyze and understand the model generated:

- The **performance of the model** with respect to a hypothetical perfect model and a random type of model,
- The **contribution of each of the explanatory variables** with respect to the target variable,
- The **significance of the various categories of each variable** with respect to the target variable.

Presentation of the SAP InfiniteInsight® User Menu

Once the model has been generated, Click *Next.* The screen *Using the Model* appears.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

The screen *Using the Model* presents the various options for using a model, that allow you to:

- Display the information relating to the model just generated or opened (*Display* section), referring to the model curve plots, contributions by variables, the various variables themselves, HTML statistical reports, table debriefing. Some information is only displayed upon request from the user: the display of InfiniteInsight® Modeler - Regression/Classification results as a decision tree, which can be specified in the modeling parameters before the model generation, or the display of model parameters, which can be requested in the general user options.
- Apply the model just generated or opened to new data, to run simulations, and to refine the model by performing automatic selection of the explanatory variables to be taken into consideration (*Run* section).
- Save the model, or generate the source code (*Save/Export* section).

Model Overview

The *Model Overview* screen displays the same information as the training summary.

Overview

Name	Name of the model, created by default from the target variable name and the data set name
Data Set	Name of the data set
Initial Number Variables	Number of explanatory variables used
Number of Selected Variables	Number of explanatory variables actually used by the resulting model.
Number of Records	Number of records in the data set
Building Date	Date and time when the model was built
Learning time	Total learning time
Engine name	Depending on the feature used: ▪ Kxen.RobustRegression ▪ Kxen.SmartSegmenter ▪ Kxen.TimeSeries ▪ Kxen.AssociationRules ▪ Kxen.EventLog ▪ Kxen.SequenceCoder ▪ Kxen.SocialNetwork

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Modeling Warnings

Monotonic Variables Detected	Indicates if monotonic variables have been found in the data set, that is, variables which direction of variation is constant, in the reading order of the data in the estimation data set.
Suspicious Variables Detected	This report presents a list of variables that are considered to be suspicious. These suspicious variables have a predictive power over 0.9, they are very correlated to the target variable. This means these variables probably bring a biased information and should not be used for the modeling. A special attention should be taken towards those variables. A more detailed report lists which variables exactly are suspicious and at which extent (see Statistical Reports > Expert Debriefing > Suspicious Variables).

Targets

For each nominal variable

<Name>	Name of the target variable
Target key	Wanted target value
Target categories Frequency	Percentage of all the target value in the Estimation data set, when dealing with a nominal target

For each continuous target variable:

<TargetName>	Name of the target variable
Min	Minimum value found for the target variable in the Estimation data set
Max	Maximum value found for the target variable in the Estimation data set
Mean	Mean of the target variable values on the Estimation data set
Standard deviation	Mean of the distance between the target values and the Mean

Performance Indicators

For each target

rr_<TargetName> *kc_<TargetName>*	Target name Note that *rr_* indicates a regression/classification and *kc_* indicates a segmentation/clustering.
Predictive Power (KI)	Quality indicator that corresponds to the proportion of information contained in the target variable that the explanatory variables are able to explain.
Prediction Confidence (KR)	Robustness indicator that signifies the capacity of the model to achieve the same performance when it is applied to a new data set exhibiting the same characteristics as the training data set.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Options

☑ To Copy the Model Summary

Click the (*Copy*) button.
The application copies the HTML code of the screen. You can paste into a word processing or spreadsheet program, a text editor, ...

☑ To Save the Model Summary

Click the (*Save*) button situated under the title.
The file is saved in HTML format.

☑ To Print the Model Summary

1. Click the (*Print*) button situated under the title.
   A dialog box appears, allowing you to select the printer to use.
2. Select the printer to use and set other print properties if need be.
3. Click *OK*.
   The report is printed.

☑ To Export to PowerPoint

1. Click the

(*Export to PowerPoint*) button.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Model Graphs

Displaying the Model Graph

☑ To Display the Model Graph

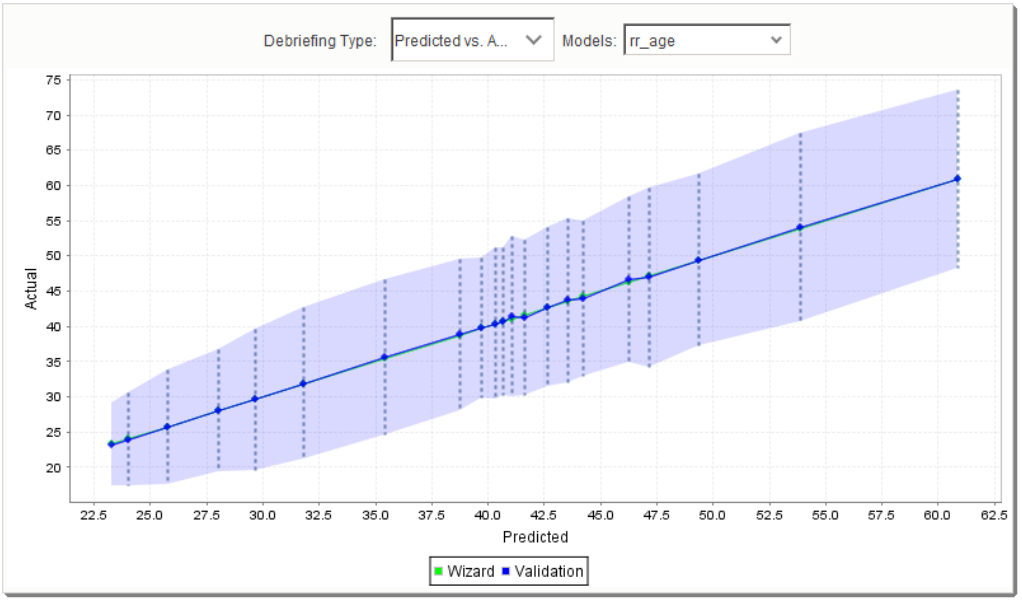
1 On the screen *Using the Model*, click the *Model Graphs* option.
The *model graphs* appears.

When the target is **nominal**, the following curve is displayed.

The default parameters display the profit curves corresponding to the *Validation* sub-set (blue line), the hypothetical perfect model (*Wizard*; green line) and a random model (*Random*; red line). The default setting for the type of profit parameter is *Detected* profit, and the values of the abscissa are provided in the form of a percentage of the entire data set.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

When the target is **continuous**, the following curve is displayed.



The default parameters display the curves corresponding to the *Validation* sub-set (blue line) and the hypothetical perfect model (*Wizard*; green line). The blue area represents the standard deviation of the current model.

For more information on the meaning of model curves (see "Understanding Model Graphs" on page 122).

2 When there is more than one target, select the target for which you want to see the curves in the *Models* list.

> **Note**
>
> To each variable corresponds a model. The name of each model is built from the *rr_* (*Robust Regression*) prefix and the model target name.

3 Select the viewing options that interest you.
For more information about viewing options.

Definition

Depending on the type of the target, the **model graph** plot allows you to:

- View the **realizable profit** that pertains to your business issue using the model generated when the target is nominal.
- **Compare the performance of the model generated** with that of a random type model and that of a hypothetical perfect model when the target is nominal.
- Compare the **predicted value** to the **actual value** when the target is continuous.

On the plot, for each type of model, the curves represent:

- when the **target** is **nominal**, the realizable profit (on the Y axis) as a function of the ratio of the observations correctly selected as targets relative to the entire initial data set (on the X axis) .
- when the **target** is **continuous**, the predicted value or score (on the X axis) in respect with the actual value or target (on the Y axis) .

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Plot Options

☑ To Display the Graphs for the Estimation, Validation and Test Sub-sets

Click *Data Set* and select one of the following options that allow you to switch between :

- the graph for the Validation sub-set ,
- the graphs for all the sub-sets .

☑ To Change the View Type

Click *View Type* and select the desired option.

☑ To Copy the Model Graph

Click the (*Copy*) button and select the desired option.
The application copies the parameters of the plot. You can paste it into a spreadsheet program (such as Excel) and use it to generate a graph.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

To Save the Model Graph

1 Click the (*Save*) button.
A dialog box appears, allowing you to select the file properties.

2 Type a name for your file.

3 Select the destination folder.

4 Click *OK*.
The plot is saved as a PNG formatted image.

To Print the Model Graph

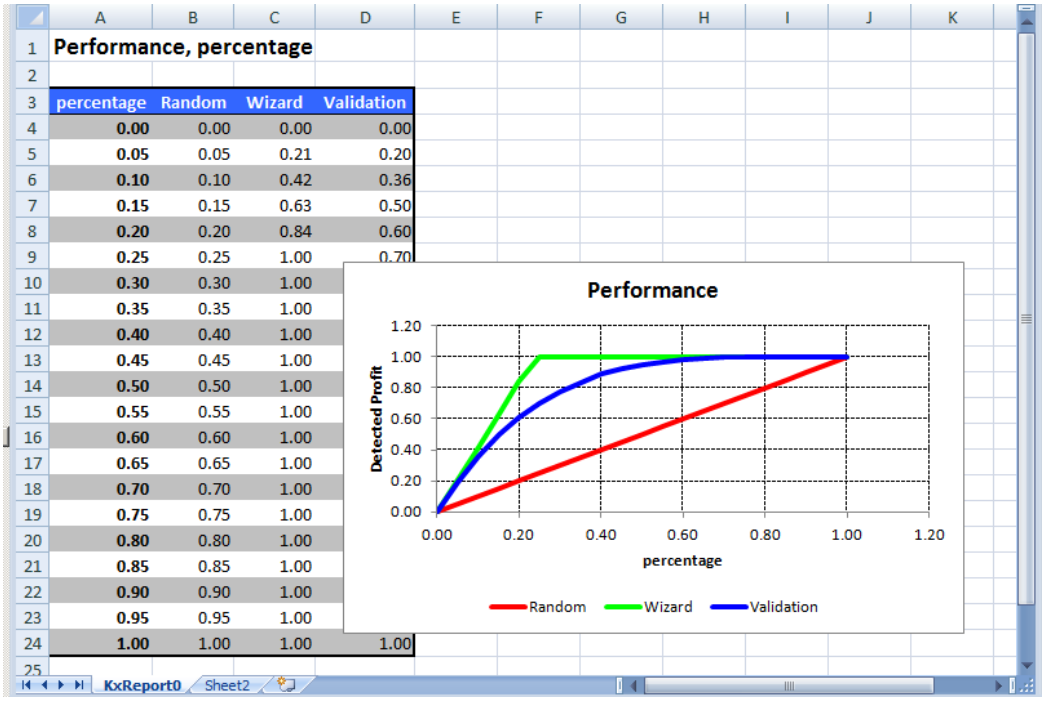
1 Click the (*Print*) button situated under the title.
A dialog box appears, allowing you to select the printer to use.

2 Select the printer to use and set other print properties if need be.

3 Click *OK*.
The report is printed.

To Export the Model Graph to Microsoft Excel

Click the (*Export to Excel*) button situated under the title. An Excel sheet opens containing the model graph you are currently viewing along with its data.



Note

compatible with Excel 2002, 2003, XP and 2007

To Open the Current Graph in a New Window

Click the (*Pin View*) button. The current graph is displayed in a new window.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

For a Nominal Target

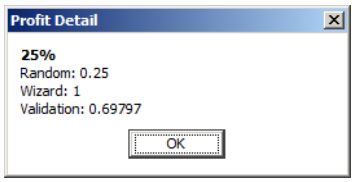
On the model curve plot, different options allow you to visualize:

- **Exact profit values** for a point for all the displayed curves.
- The curves for the different profit types: Detected, Lift, Normalized, and Customized.

For more information on profit types (see "Available Profit Types" on page 46).

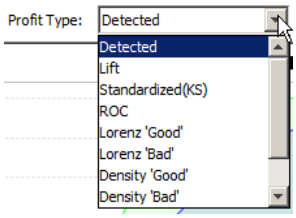
☑ To Display the Exact Profit Values for a Given Point

On the screen *Model Curves*, on the plot, click a point on one of the curves presented.
For instance, by clicking a point on any one of the curves whose value on the abscissa is 25%, the exact profit values appear.



☑ To Select a Profit Type

1 On the screen *Model Curves*, beneath the plot, click the drop-down list associated with the *Profit* field.
The list of profit types appears.



2 Select a profit type.
The corresponding profit curves appear.

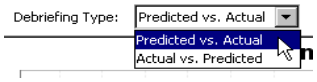
For a Continuous Target

☑ To Display the Exact Profit Values for a Given Point

On the screen *Model Graphs*, on the plot, click a point on one of the curves presented.

☑ To Select the Debriefing Type

1 On the screen *Model Graphs*, above the plot, click the drop-down list associated with the *Debriefing Type* field.
The list of debriefing types appears.



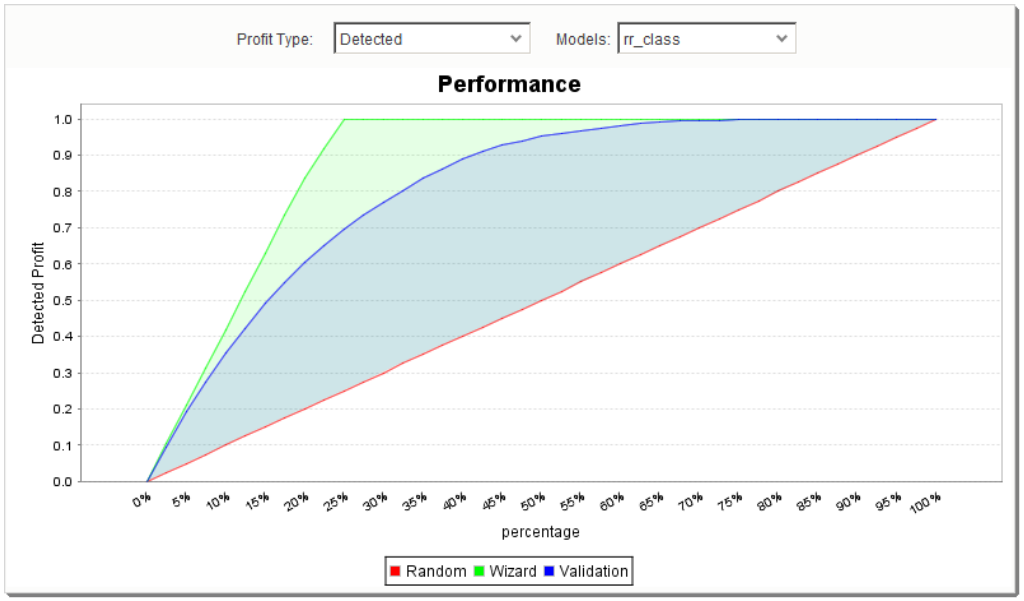
2 Select a debriefing type.
The corresponding plot appears.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Understanding Model Graphs

For a Model with a Nominal Target

The following figure represents the model graph produced using the default parameters.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

On the plot, the curves for each type of model represent the profit that may be realized (Y axis), that is, the percentage of observations that belong to the target variable, in relation to the number of observations selected from the entire initial data set (X axis). On the X axis , the observations are sorted in terms of decreasing "score", that is, the decreasing probability that they belong to the target category of the target variable.

In the application scenario, the model curves represent the ratio of prospects likely to respond in a positive manner to your marketing campaign, relative to the entire set of prospects contained in your database.

Detected profit is the default setting for type of profit. Using this type of profit:

- The value "0" is assigned to observations that do not belong to the target category of the target variable,
- The value "1/(frequency of the target variable in the data set)" is assigned to observations that do belong to the target category of the target variable.

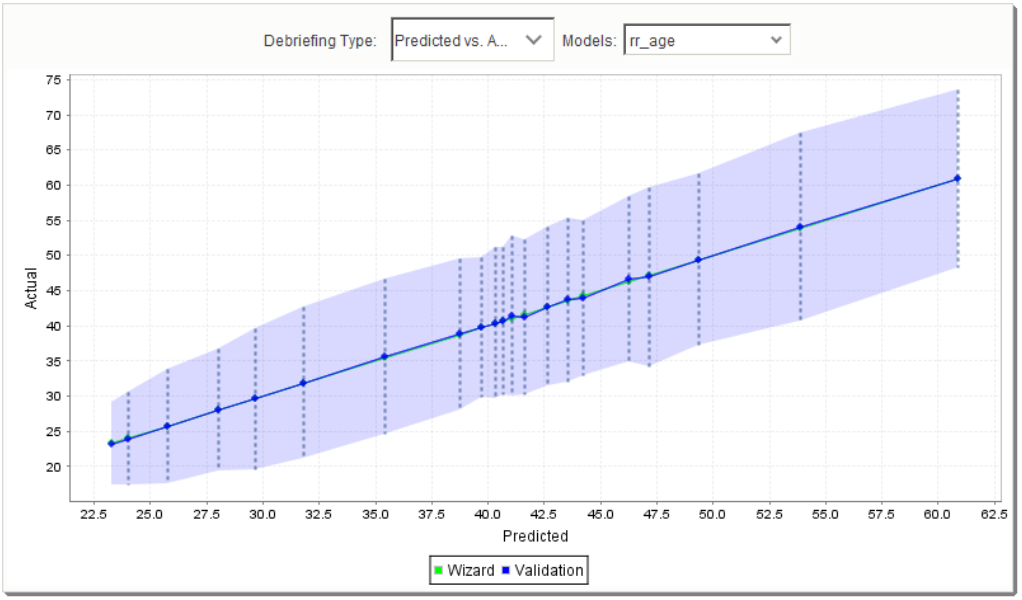
The following table describes the three curves represented on the plot created using the default parameters.

The curve...	Represents...	For instance, by selecting...
Wizard (green curve, at the top)	The profit that may be achieved using the **hypothetical perfect model**, that allows one to know with absolute confidence the value of the target variable for each observation of the data set	25% of the observations from your entire data set with the help of a perfect model, 100% of observations belonging to the target category of the target variable are selected. Thus maximum profit is achieved. **Note** These 25% correspond to the proportion of prospects who responded in a positive manner to your marketing campaign, during your test phase. For these prospects, the value of the target variable, or profit, is equal to 1.
Validation (blue curve, in the middle)	The profit that may be achieved using **the model generated by InfiniteInsight® Modeler - Regression/Classification**, that allows one to perform the best possible prediction of the value of the target variable for each observation of the data set	25% of the observations from your initial data set with the help of the model generated, 66.9% of the observations belonging to the target category of the target variable are selected.
Random (red curve, at the bottom)	The profit that may be achieved using a **random model**, that does not allow one to know even a single value of the target variable for each observation of the data set	25% of the initial data set using a random model, 25% belonging to the target category of the target variable are selected.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

For a Model with a Continuous Target

The following graph represents the model curve plot produced using a continuous target.



The default graphic displays the *actual target* values as a function of *predicted target* values. Two curves are displayed: one for the *Validation* sub-set (blue line) and another for the hypothetical perfect model (*Wizard*; green line). The *Validation* curve gives *Actual Target* value as a function of *Predicted Target* value. For example, when the model predicts 35, the average actual value is 37. The *Wizard* curve is just X=Y, meaning that all the predicted values are equal to the actual values. The graph is an easy way to quickly see model error. When the curve is going far from Wizard, it means that the predicted value is suspicious.

The graph is computed as follow:

- about 20 segments or bins of predicted values are built. Each of these segments represents roughly 5% of the population.
- for each of these segments, some basic statistics are computed on actual value, such as the mean of the segment **(SegmentMean)**, the mean of the associated target **(TargetMean)** and the variance of this target within that segment **(TargetVariance)**. For example for predicted value in [17; 19], the mean would be 18.5, the actual target mean would be 20.5 and the actual target variance would be 9. In this case we could say that, if the predicted value is between 17 and 19, the model is underestimating a bit the actual value.

For each curve, a dot on the graph corresponds to the segment mean on the X-axis, and the target mean on the Y-axis.

The blue area represents the expected deviation of the current model. The blue area shows where about 70% of the actual values are expected to be. In other words, it means that, in case of a Gaussian distribution, about 70% of the actual points should be in the blue area (keep in mind that this is a theoretical percentage that may not be observed every time). The default setting for the type of curve parameter is *Predicted vs Actual*. The extreme values for prediction ranges are `{TargetMean - (sqrt(TargetVariance)); TargetMean + (sqrt(TargetVariance))}`.

> **Note**
>
> `sqrt(TargetVariance)` is equal to the Standard Deviation.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Predictive Power, Prediction Confidence and Model Graphs

On the model graph plot:

- Of the estimation data set (default plot), the **predictive power** corresponds to "the area found between the curve of the **model generated** and that of the **random model**" divided by "the area found between the curve of the **perfect model** and that of the **random model**". As the curve of the generated model approaches the curve of the perfect model, the value of the predictive power approaches 1.
- Of the estimation, validation and test data sets (select the corresponding option from the list *Data set*, located below the plot), the prediction confidence corresponds to one minus "the area found between the curve of the **estimation data set** and that of the **validation data set**" divided by "the area found between the curve of the **perfect model** and that of the **random model**".

Contributions by Variables

Definition

The *Contributions by Variables* plot allows you to examine the relative significance of each of the variables within the model. On this plot, each bar represents the contribution of an explanatory variable with respect to the target variable.

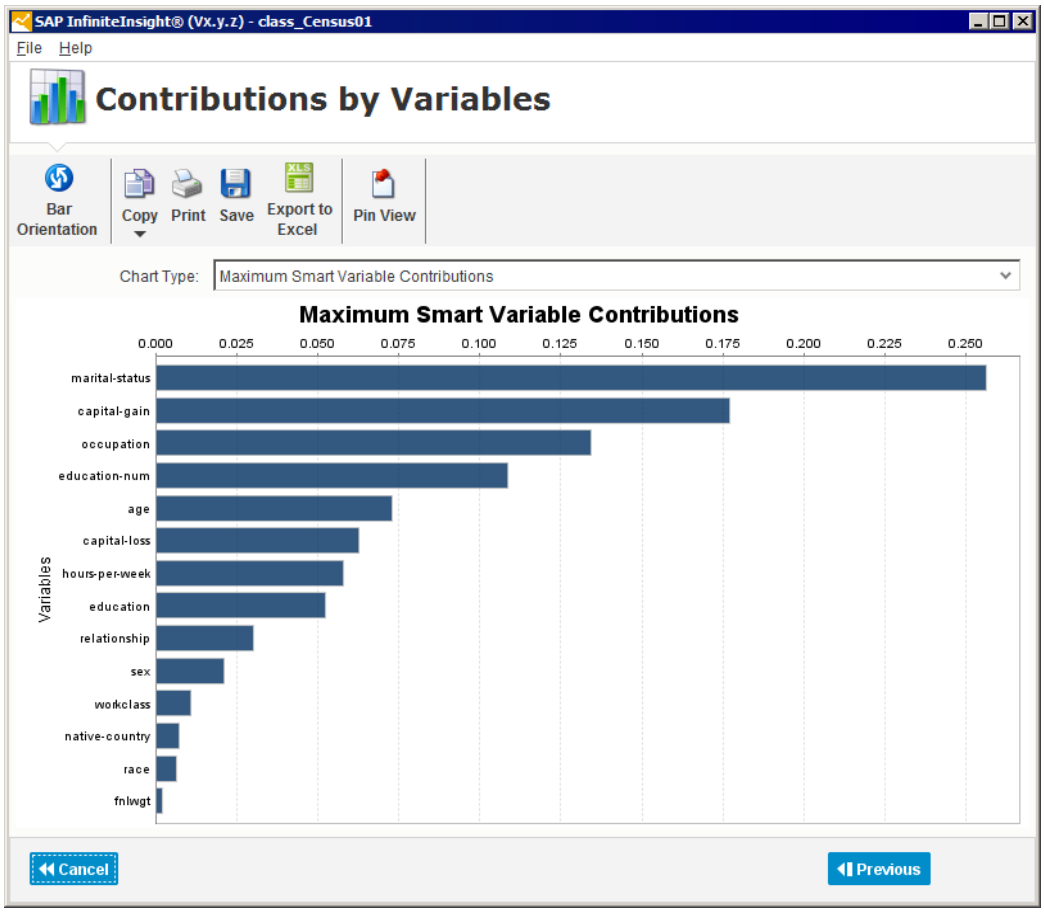
The following four types of plots allow you to visualize contributions by variables:

- *Variable Contributions*, that is, relative importance of each variable in the built model.
- *Variable Weights*, that is, weights (in the final polynomial) of the normalized variables.
- *Smart Variable Contributions*, that is, the variables internal contributions.
- *Maximum Smart Variable Contributions*, that is, the maximum smart variable contributions including only the maximum of similar variables. For example, only binned encoding of the continuous variable *age* will be displayed. This is the chart displayed by default.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Display the Plot of Contributions by Variables

1 On the screen *Using the Model*, click the option *Contributions by Variables*.
The plot *Contributions by Variables* appears.
The default plot type is *Maximum Smart Variable Contributions*.



If your data set contains date or datetime variables, automatically generated variables can appear in this panel. For more information, refer to section Date and Datetime Variables: Automatically Generated Variables (on page 30).

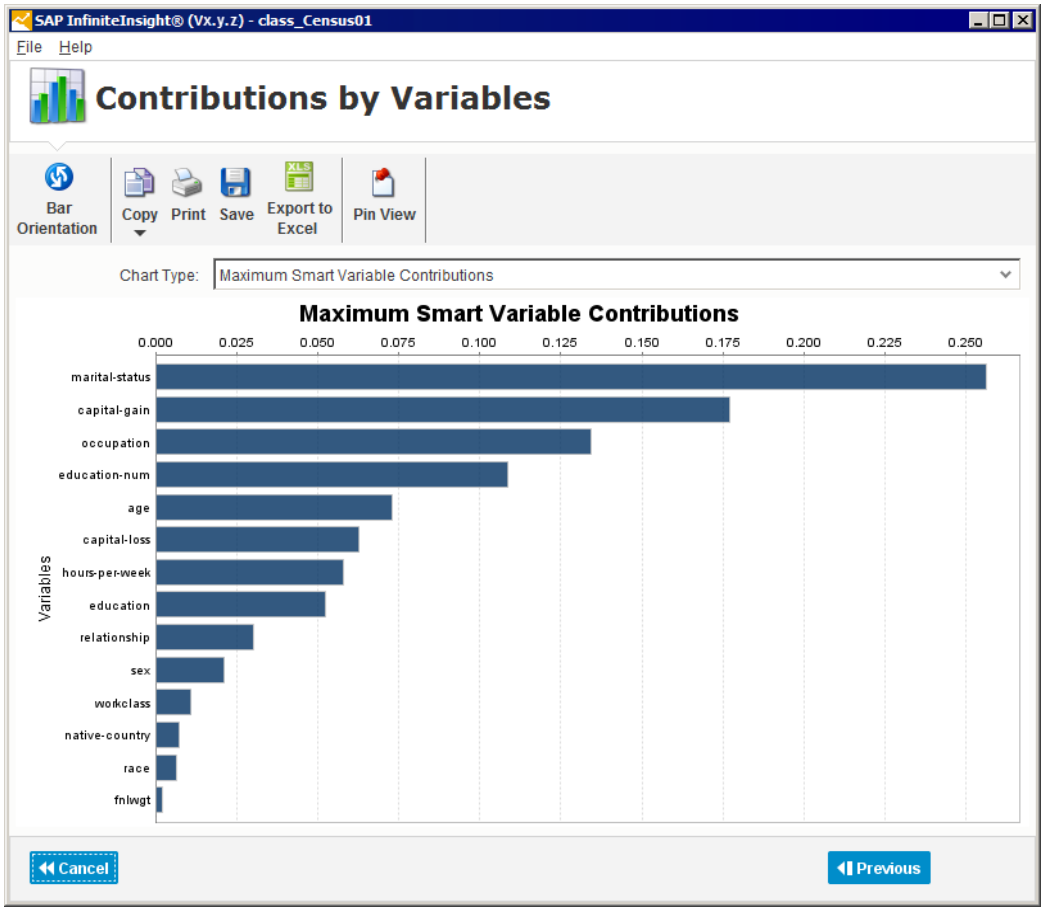
2 You can drill down on a variable, that is, display the plot of details of this variable, where the categories of the variable can be seen. To zoom in on a variable, double-click the corresponding bar. Go to section Significance of Categories.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Understanding Contributions by Variables

Only the plot, *Maximum Smart Contributions by Variables*, the default selection, is presented in this guide.

The *Contributions by Variables* option allows the user to examine the relative significance of each of the explanatory variables in relation to the target variable. This significance is relative, as the weight of each variable is pro-rated as a function of the significance of the other explanatory variables.



The plot above corresponds to the model generated, and illustrates the two variables that contribute the most to the target variable, which, in this scenario, are:

- *marital-status*,
- *capital-gain*.

In other words, the *marital-status* and *capital-gain* variables are those which have the greatest effect on whether a prospect will respond positively or negatively to your marketing campaign. Among all the variables included in the sample data set, these two are the most discriminatory variables with respect to the target variable *Class*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Correlated Variables

To say that variables are **correlated** implies a certain level of redundancy, that they each contribute some of the same information with respect to the target variable. Two variables said to be highly correlated would describe the same information, or the same concept, to an even greater degree.

The plot *Smart Variable Contributions* reflects the correlation that may exist between various explanatory variables. When two variables A and B are strongly correlated:

- Variable A, with a greater contribution than B with respect to the target variable, becomes the "primary variable": the plot displays **all its information** including what it has in common with variable B.
- Variable B, with a smaller contribution than A with respect to the target variable, becomes the "secondary variable": only its **marginal contribution** is displayed on the plot, meaning that only the supplementary contribution to target variable information, or the values that B does not share with A, are displayed. This difference of information is noted `[variable_B]-[variable_A]`.

Encoded Variables

Creating an SAP InfiniteInsight® model uses not only the original variables but also, in case of continuous or ordinal variables, their value as encoded by InfiniteInsight® Modeler - Data Encoding. This is called **dual-encoding** and allows SAP InfiniteInsight® to find all the information contained in each variable.

The encoded variables appear on the variable contributions plots with the prefix `c_`. For example, the encoded version of a continuous variable named `age` is noted `c_age`.

Note

In InfiniteInsight® Modeler, on the Data Description panel, if you enable the Natural Encoding for a given variable, its K2C-encoded value (*c_variableName*) will not be generated.

Category Significance

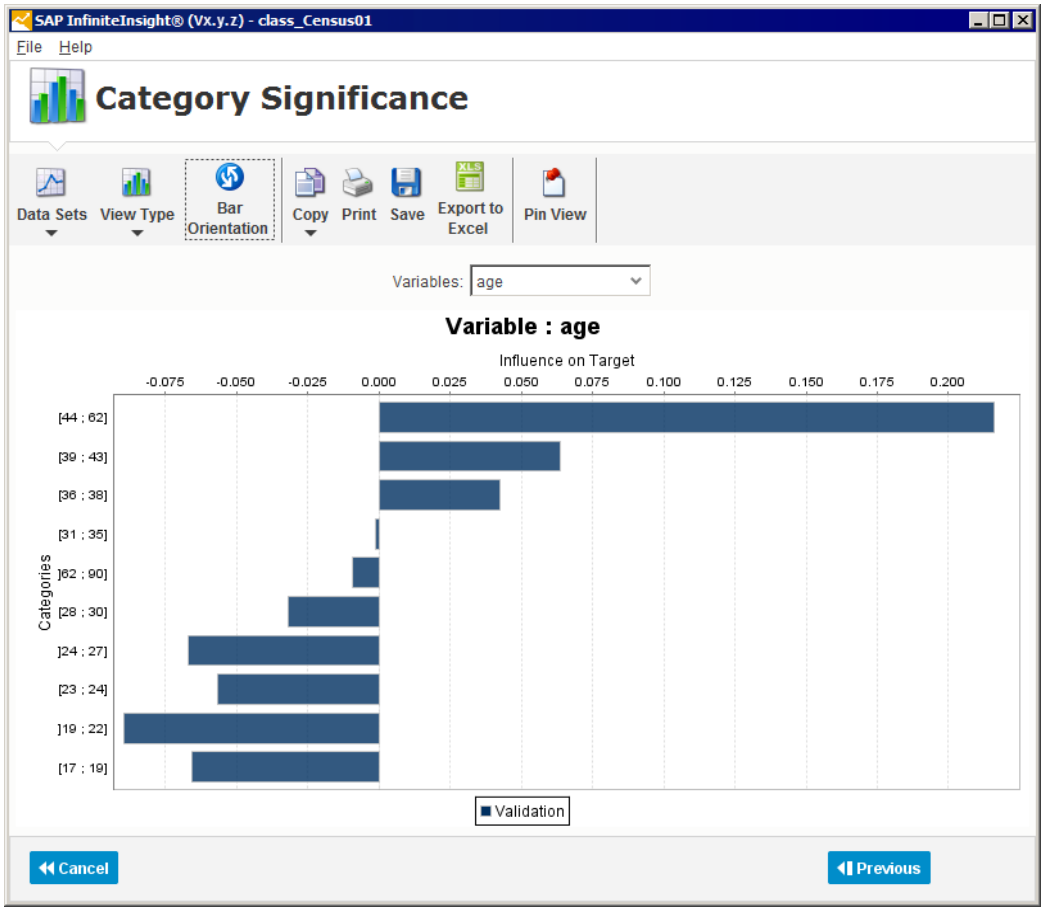
Definition

The *Significance of Categories* plot illustrates the relative significance of the different categories of a given variable with respect to the target variable.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Display the Significance of Categories Plot

1 On the screen *Using the Model*, click *Category Significance*.
The plot *Category Significance* appears.



2 In the *Variables* list located above the plot, select the variable for which you want to display the categories.

If your data set contains date or datetime variables, automatically generated variables can appear in the *Variables* list. For more information, refer to section Date and Datetime Variables: Automatically Generated Variables (on page 30).

Note

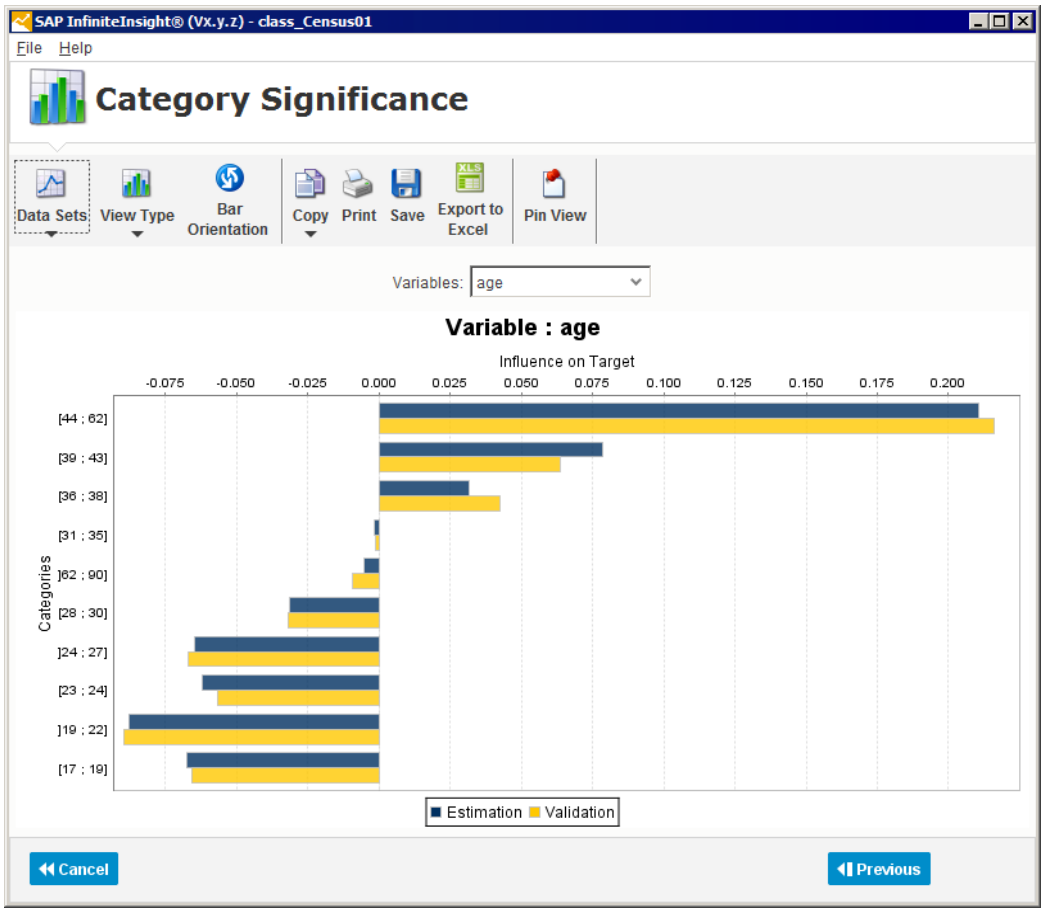
- You can display the relative significance of the categories of a variable directly from the plot *Contributions by Variables*. On the plot *Contributions by Variables*, double-click the bar of the variable which interests you.

- In case no user structure has been defined for a continuous variable, the plot *category significance* displays the categories created automatically using the band count parameter. The number of categories displayed corresponds to the value of the band count parameter. For more information about configuring this parameter, please refer to the section Band Count for Continuous Variables.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Switch Between "Validation Data Set" and "All Data Sets" Plots

1 Click *Data Sets* and select the (*All Data Sets*) button to display all data sets.
The plot displaying all data sets appears.

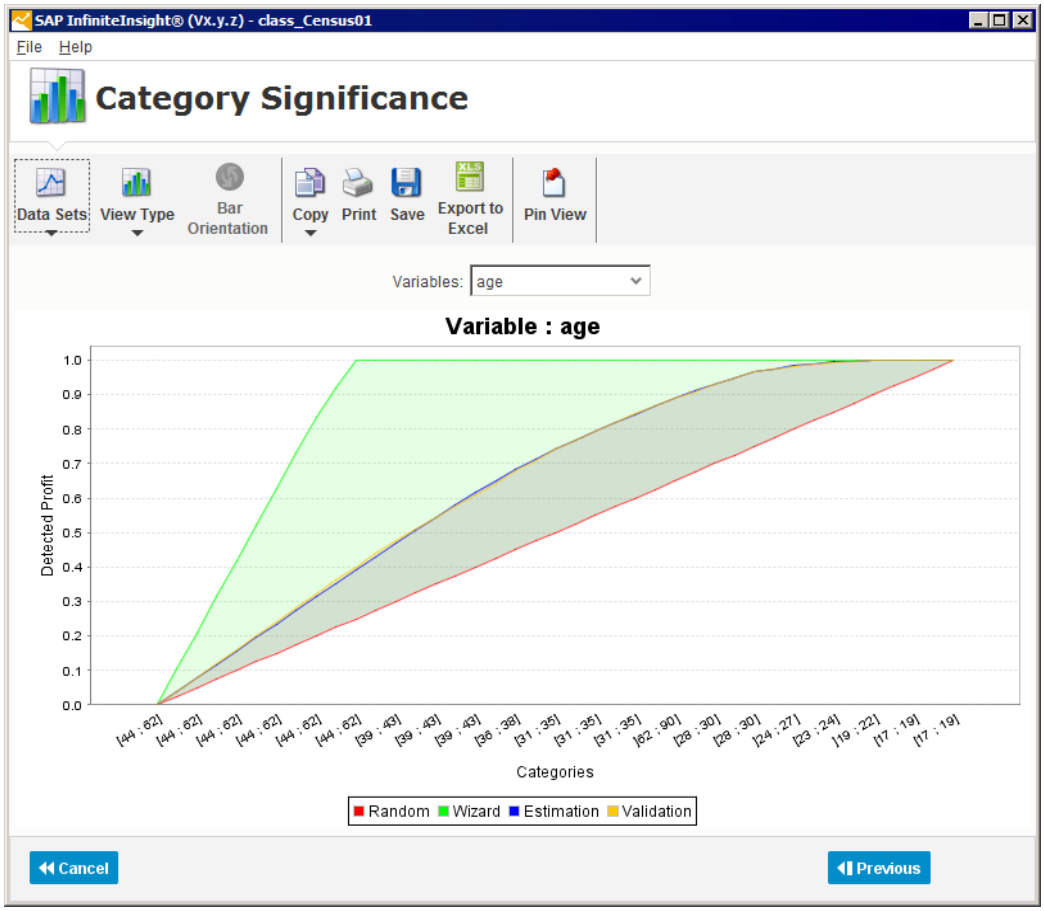


© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2 Click *Data Sets* and select the (*Validation Only*) button to go back to the *Validation Data Set* plot.

☑ To Switch between Curve and Bar Charts

1 Click *View Type* and select the button to display the curve chart.
The curve plot appears.



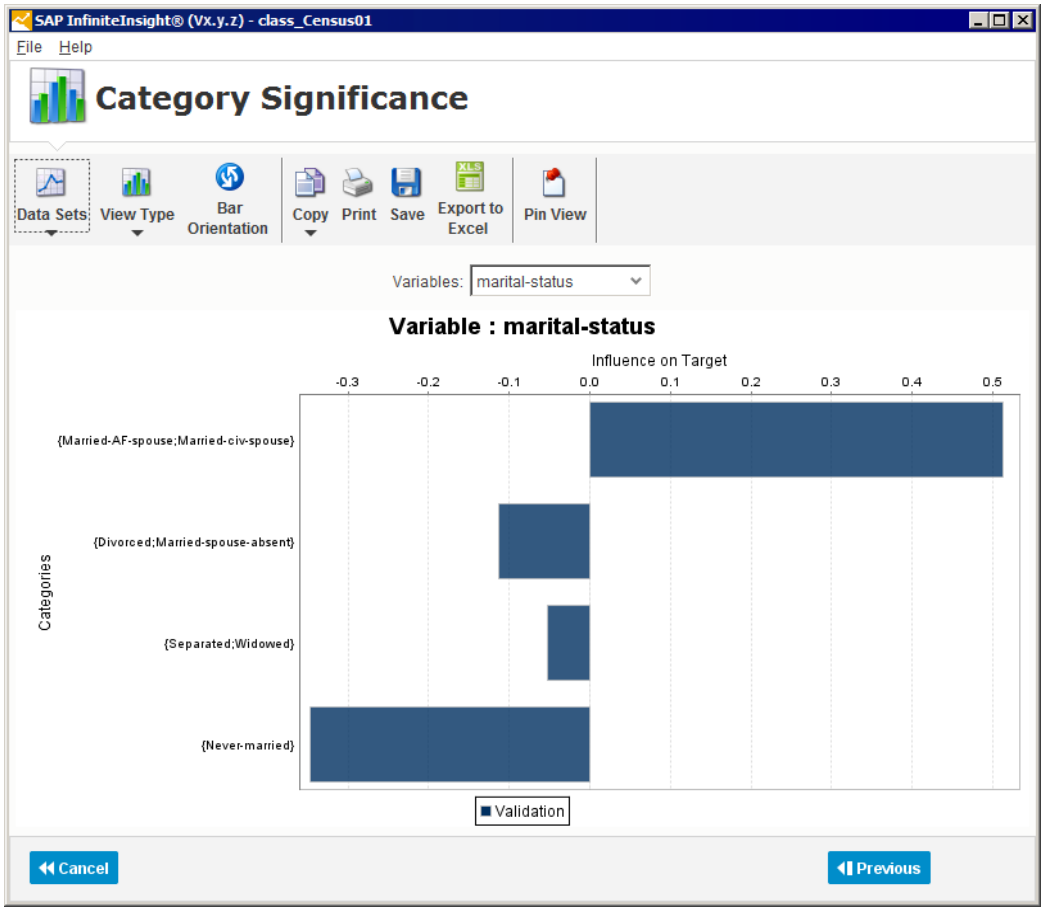
2 Click *View Type* and select the button to go back to the bar chart.

> **Note**
>
> You can combine the different types of plot. For example, you can display *All Datasets* in a curve chart or the *Validation Data Set* in a bar chart.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

For this Scenario

Select the variable *marital-status*, which is the explanatory variable that contributes the most to the target variable *Class*.



This plot presents the effect of the categories of the *marital-status* variable on the target variable.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Variable Categories and Profit

The plot *Category Significance* illustrates the relative significance of the different categories of a given variable with respect to the target variable.

On this type of plot:

- **The higher on the screen one finds a category**, the greater the positive effect on the target category (or hoped-for value) of the target variable. In other words, the higher a category appears on the screen, the more representative that category is of the target category of the target variable.
- **The width and direction of the bar correspond to the profit contributed by that category**. In other words, they correspond to the relationship of that category to the target variable and whether that category has more or less observations belonging to the target category of the target variable. For a given category, a **positive bar** (on the right of 0.0) indicates that the category contains more observations belonging to the target category of the target variable than the mean (calculated on the entire data set). A **negative bar** (on the left of 0.0) indicates that the category contains a lower concentration of target category of the target variable than the mean.

> **Note**
>
> You can display the profit curve for the selected variable by clicking the button (*Display Profit Curve*) located in the tool bar under the title.

The importance of a category depends on both its difference to the target category mean and the number of represented cases. High importance can result from:

- a high discrepancy between the category and the mean of the target category of the target variable
- or a minor discrepancy combined with a large number of records in the category
- or a combination of both.

The width of the bar shows the profit from that category. The positive bars correspond to categories which have more than the mean number from the target category (that is, responders) and the negative bars correspond to categories which have less than the mean number from the target category (that is, responders).

The **Variables** pull down menu allows the selection and graphing of any of the variables in the model.

The tool bar located under the title allows the user to copy the coordinates to the clipboard, print the plot or save it in PNG format.

The values are normalized and their sum always equals to "0". Depending on the chosen profit strategy, or on the continuous target variables value type, you can obtain all positive importances or negative and positive importances.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Axes

The **X-axis** shows the influence of the variable categories on the target. The significance of the different numbers on the X-axis are detailed in the following table.

Number on the X-axis	Indicates that the category has...
positive number	a positive influence on the target
0	no influence on the target (the behavior is the same as the average behavior of the whole population)
negative number	a negative influence on the target

The **Y-axis** displays the **variable categories**. Categories sharing the same effect on the target variable are grouped. They appear as follow `[Category_a;Category_b;Category_c]`. Categories not containing sufficient numbers to provide robust information are grouped in the `KxOther` category. When a variable is associated with too many missing values, the missing values are grouped in the `KxMissing` category. Both categories are created automatically by SAP InfiniteInsight®.

Category Importance Definition

The following definition applies to continuous targets; some wording may be simplified for binary targets. The formulas presented below can also be applied to the binary target case (use categories instead of segments in this case).

We consider the case where a InfiniteInsight® Modeler - Regression/Classification regression model is trained on a continuous target/signal S with the help of an input variable X.

InfiniteInsight® Modeler - Regression/Classification starts by binning the continuous target S into B segments: $S_{1,...,}S_B$ and by computing the basic statistics of the inputs and the cross statistics of the inputs with respect to the target.

We will suppose that the input X is a nominal (categorical) variable, though the whole process can be extended easily to the case of ordinal and continuous inputs.

We will suppose that X has N categories: $X_1, ..., X_N$.

We are interested in assessing the importance of a category Xi with respect to the target S.

The importance of a category depends on two factors:

- The fact that the distribution of the target for this category is significantly skewed towards high values or low values when compared with the distribution of the target on the entire population
- The frequency of this category.

High importance can result from either of the following:

- a high discrepancy between the target distribution for cases associated to this category and the distribution of the target variable for the entire population
- a minor discrepancy combined with a large number of records in the category
- a combination of both

SAP InfiniteInsight® uses a non-parametric setting in which the category importance is defined as:

$$CategoryImportance(Xi) = normalProfit(Xi) * \frac{Freq(Xi)}{Z}$$

where :

- $normalProfit(Xi)$ is the normal profit of category Xi, see below for a definition
- $Freq\ (Xi)$ is the global frequency of the category Xi
- Z is a normalization constant.

We give below the details of the computation of these quantities.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Normal Profit

Each category of the target S_j is associated with a profit $profit(Sj)$ defined such that:

$$\sum_{j=1}^{j=B} profit(S_j)\ Freq(Sj) = 0$$

The profit of a target category is a value in the range $[-1;\ +1]$. It is defined the following way from the (cumulated) target category frequencies:

$$profit(S_j) = 2\sum_{k=1}^{j-1} Freq(S_k) + Freq(S_j) - 1$$

The normal profit of a category Xi is then defined as:

$$normalProfit(X_i) = \sum_{j=1}^{j=B} profit(S_j)\ Prob[S_j \mid X_i]$$

Where $Proba[Sj\ |Xi]$ is the conditional probability of observing the target category Sj in the variable category Xi (cross statistics):

$$Prob[S_j \mid X_i] = \frac{Freq(S_j\ ;\ X_i)}{Freq(X_i)}$$

The fact that these formulas rely only on frequencies makes them resistant to any monotonic transformation of the target S.

Normalization Constant

The normalization can be approximated for non pathological continuous targets (that is continuous targets without distribution peak (Dirac)) as:

$$Z = P[S > median(S)] * (1 - P[S > median(S)])$$

In most cases, a good approximation is = 0.25 .

 © 2014 SAP AG or an SAP affiliate company. All rights reserved-

Normal Profit Properties

There are several interesting things to note about normal profit:

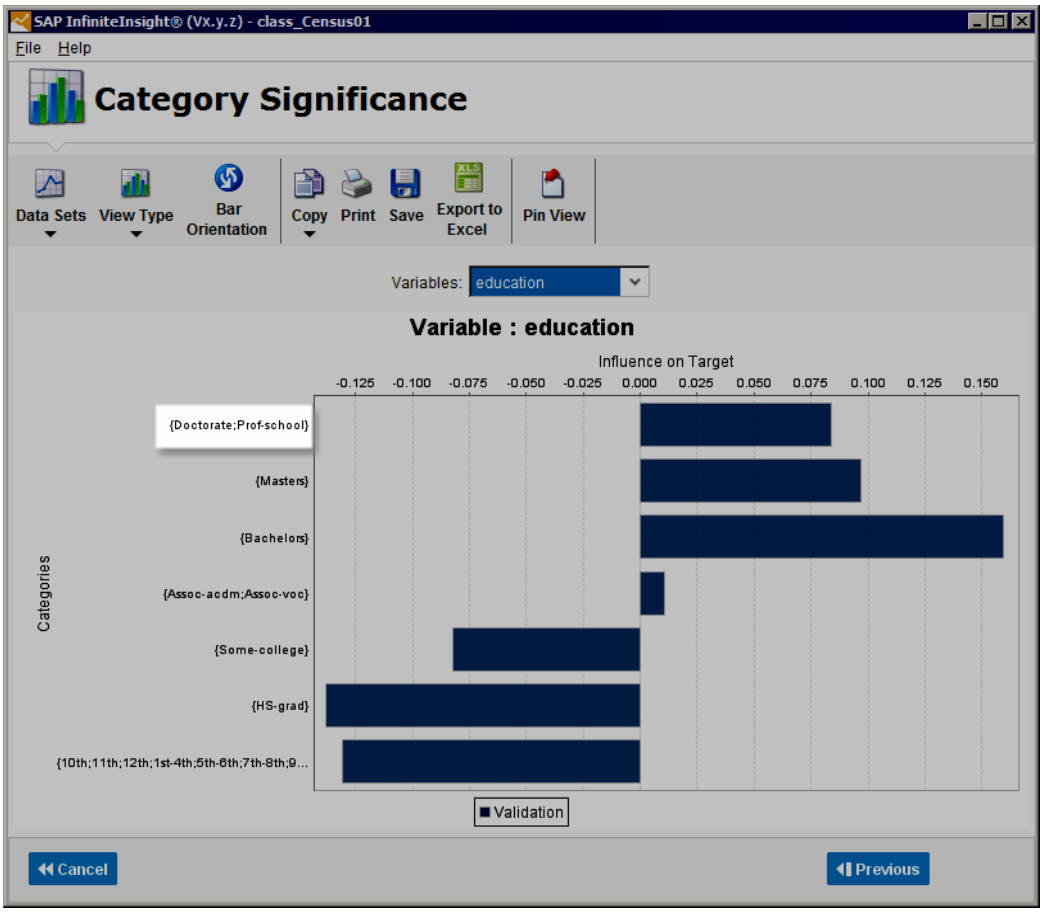
1. The normal profit of category is independent of the target values themselves (user can change the target value through monotonic transformations; the normal profit of the categories with respect to this target will not change). This belongs to non-parametric metrics.
2. A consequence of 1 is that this metric is resistant to outliers: when there are a few occurrences of the target with very high values with respect to the rest of the target value distributions, the notion of normal profit is not impacted.
3. The weighted sum of the normal profit for all categories of a given variables will always be 0.

Grouping Categories

On the plot of details of a variable, categories may appear grouped.

When the option *Optimal Grouping* is enabled, SAP InfiniteInsight® groups those **categories sharing the same effect on the target variable**. For instance, for the variable *education*, the categories *Doctorate* and *Prof-School* are grouped. If the explanatory variable is continuous, SAP InfiniteInsight® identifies the points where behavioral changes occur with respect to the target variable and automatically crops the variable into intervals exhibiting homogeneous behavior with respect to the target.

For more information, please see section Optimal Grouping for all Variables.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

When **categories do not contain sufficient numbers to provide robust information**, they are grouped in the *KxOther* category, that is created automatically.

When a variable is associated with **too many missing values**, the missing values are grouped in the *KxMissing* category, that is also created automatically.

To understand the value of the categories *KxOther* and *KxMissing*, consider the following example. The database of corporate customers of a business contains the variable "web address". This variable contains the Web site address of the corporate customers contained in the database. Some companies have a Web site; others do not. In addition, each Web site address is unique. In this case, SAP InfiniteInsight® automatically transforms the "web address" variable into a binary variable with two possible values: *KxOther* (the firm has a Web site) and *KxMissing* (the firm does not have a Web site).

Statistical Reports Options

A tool bar is provided allowing you to modify how the current report is displayed, to copy the report, to print it, to save it or to export it to Excel.

Statistical Reports

The *Statistical Reports* provide you with a set of tables that allows you a more detailed debriefing of your model. These reports are grouped in different levels of debriefing:

- the *Descriptive Statistics*, which provides the statistics on the variables, their categories and the data sets, as well as the variables cross-statistics with the target

> **Note**
>
> If your data set contains date or datetime variables, automatically generated variables will appear in the statistical reports. For more information, refer to section Date and Datetime Variables: Automatically Generated Variables (on page 30).
>
> In the section *Cross Statistics with the Target(s)*, the number of displayed categories corresponds to:
>
> - the number of categories as defined in the user structure
> - the band count if no user structure has been defined.
>
> For more information, see the section Band Count for Continuous Variables.
>
> In the section *Grouped Cross Statistics with the Target(s)*, if the option *Optimal Grouping* is enabled, the number of displayed categories is lower than that defined:
>
> - in the user structure
> - by the parameter *band count* if no user structure has been defined.
>
> For more information about using this parameter, refer to the section Optimal Grouping for all Variables.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

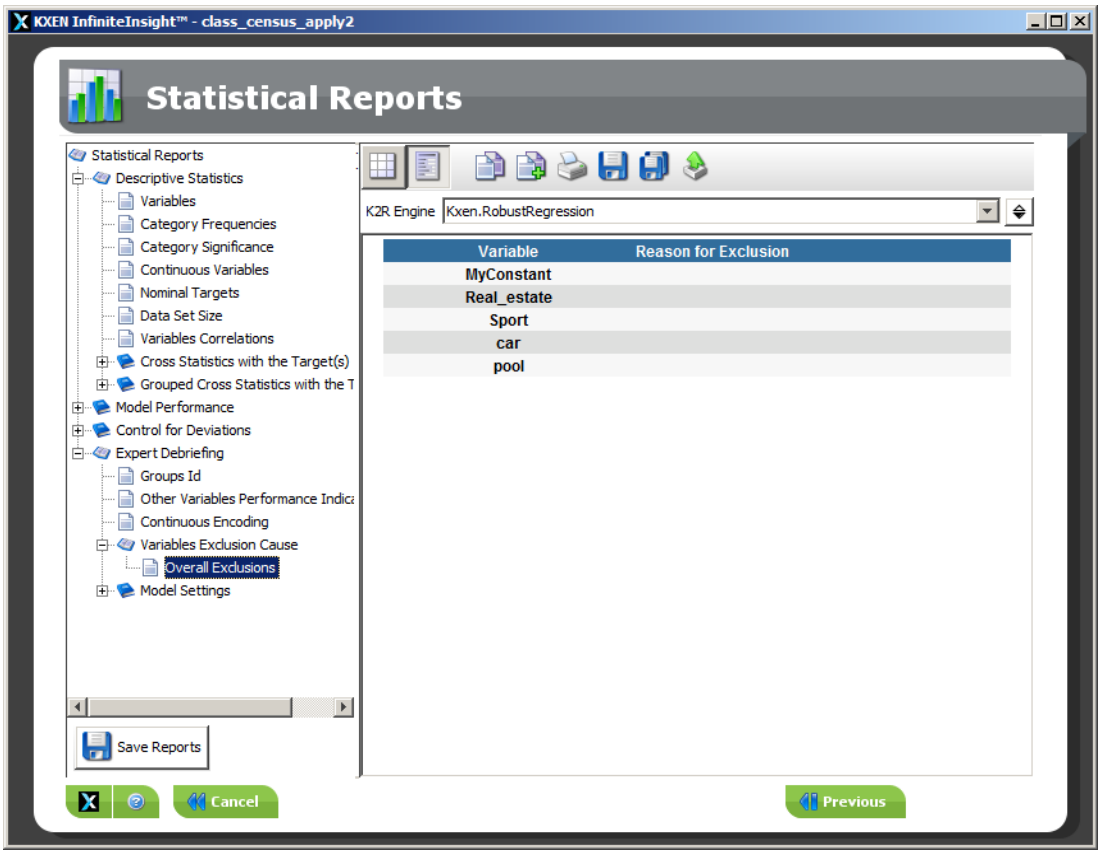
- the *Model Performance*, in which you will find the model performance indicators, the variables contributions and the score detailed statistics.
- the *Control for Deviations*, which allows you to check the deviations for each variable and each variable category between the validation and test data sets.
- the *Expert Debriefing*, in which you will find more specialized performance indicators, as well as the variables encoding, the excluded variables during model generation and the reason for exclusion, and so on.

Variables Exclusion Cause

Statistical Reports include the section *Variables Exclusion Causes*. For regression and classification models, this sections presents the reason why a variable got excluded from the model, if any variable was.

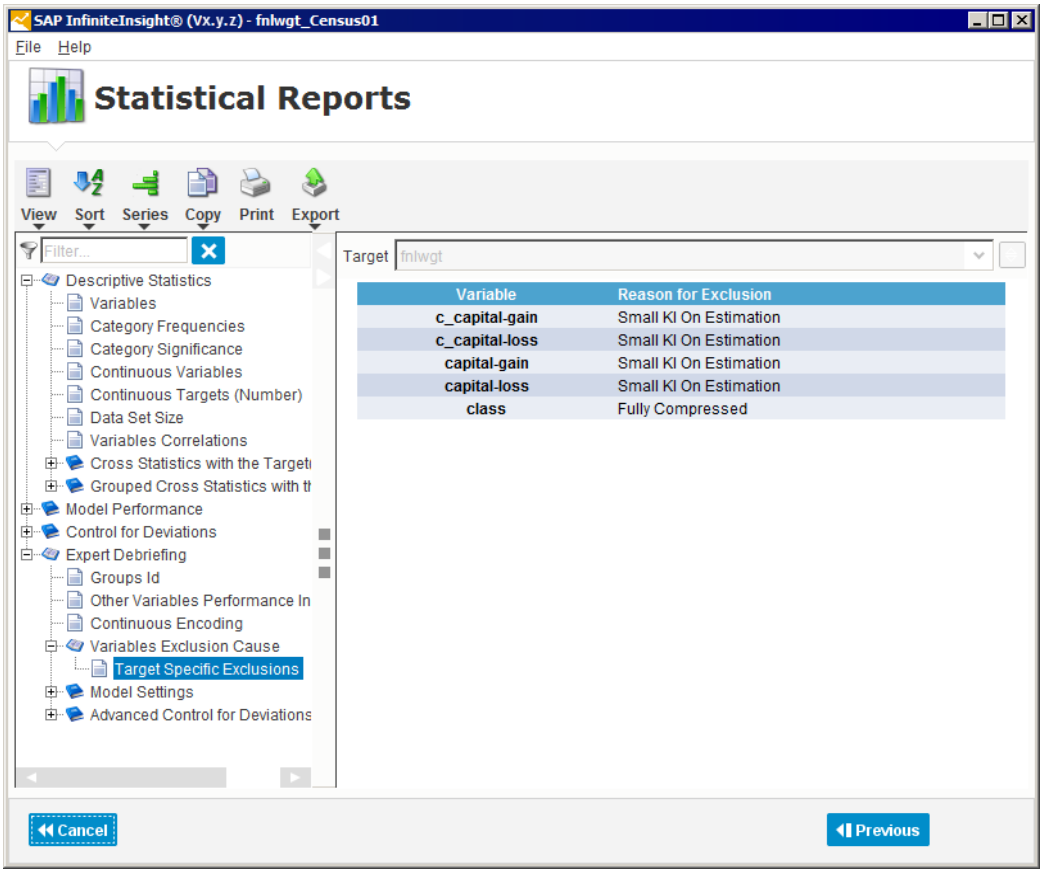
It is divided in two parts:

- *Overall Exclusions*, showing the variables excluded from the whole model



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

- *Target Specific Exclusions*, showing the variables excluded towards a particular target.



 © 2014 SAP AG or an SAP affiliate company. All rights reserved-

Possible Variables Exclusion Causes

The table below shows the possible variables exclusion causes:

Overall Exclusions / Target Specific Exclusions	Name	Explanation
Overall Exclusions	Constant	The variable has only one value (continuous variables) or one category (nominal or ordinal variables) in the data set. The variable is discarded with respect to all targets.
Overall Exclusions	Small Variance	For continuous variables, the variance is small. The variable variation is noise. The variable is discarded with respect to all targets.
Target Specific Exclusions	Fully Compressed	The variable has been fully compressed with respect to the target. It will be excluded from the model with respect to this target.
Target Specific Exclusions	Small KI on Estimation	The variable has a small KI on Estimation data set with respect to the target. It will be excluded from the model with respect to this target.
Target Specific Exclusions	Small KI on Validation	The variable has a small KI on Validation data set with respect to the target. It will be excluded from the model with respect to this target.
Target Specific Exclusions	Large KI Difference	A large KI difference has been observed for this variable between Estimation and Validation data sets with respect to the target. It will be excluded from the model with respect to this target.
Target Specific Exclusions	Small KR	The variable has a small KR with respect to the target. It will be excluded from the model with respect to this target.

Display Options

View		This option allows you to display the current report view in the graphical table that can be sorted by column.
		This option allows you to display the current report view as a HTML table.
		Some reports can be displayed as a bar chart. This bar chart can be sorted by ascending or descending values, or by ascending or descending alphabetical order. You can also select which data should be displayed.
		Some reports can be displayed as a pie chart.
		Some reports can be displayed as a line chart.
Sort		When the current report is displayed as a bar chart, this option allows you to change the orientation of the bars (from horizontal to vertical and vice versa).
		This option allows you to display the current report with no sorting.
		This option allows you to sort the current report by ascending values.
		This option allows you to sort the current report by descending values.
		This option allows you to sort the current report by ascending names.
		This option allows you to sort the current report by descending names.

Series	This option allows you to select which columns to display for current report.

Usage Options

		This option allows you to copy the data from the current view of the displayed report. The data can then be pasted in a text editor, a spreadsheet, a word processing software.
		This option allows you to print the current view of the selected report depending on the chosen display mode (HTML table, graph, ...).
		This option allows you to save under different formats (text, html, pdf, rtf) the data from the current view of the selected report.
		This option allows you to save under different formats (text, html, pdf, rtf) the data from all the views of the selected report.
		This option allows you to export to Excel.
		This option allows you to save all reports.
		This option allows you to save the customized style sheet.

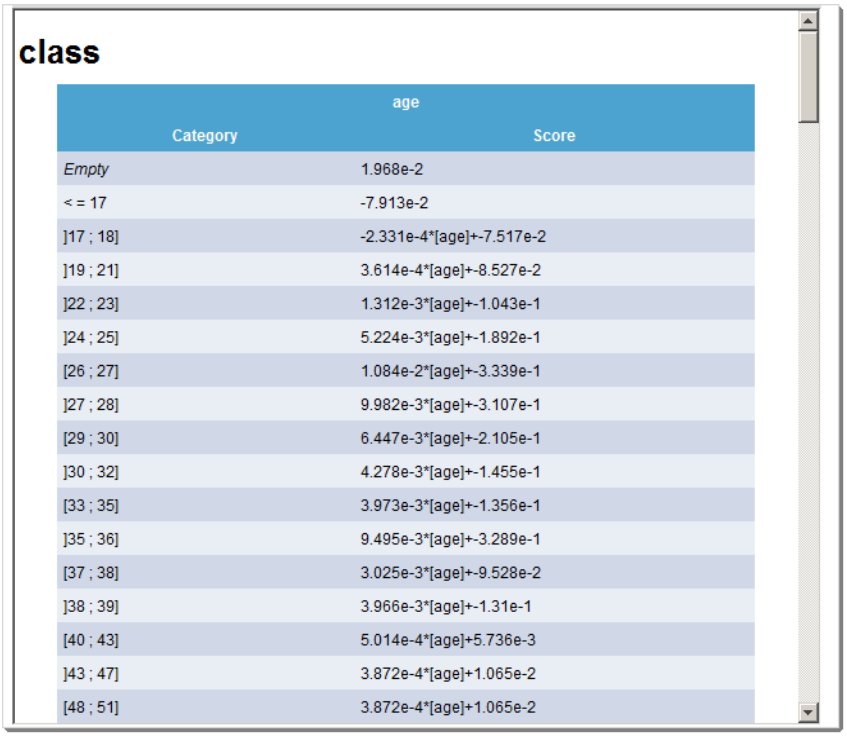
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Scorecard

This screen provides you with the coefficients associated to each category for all variables in a **regression model**.

☑ To Obtain a Score

Add all the coefficients corresponding to the selected value of each variable.



class

age	
Category	Score
Empty	1.968e-2
< = 17	-7.913e-2
]17 ; 18]	-2.331e-4*[age]+-7.517e-2
]19 ; 21]	3.614e-4*[age]+-8.527e-2
]22 ; 23]	1.312e-3*[age]+-1.043e-1
]24 ; 25]	5.224e-3*[age]+-1.892e-1
[26 ; 27]	1.084e-2*[age]+-3.339e-1
]27 ; 28]	9.982e-3*[age]+-3.107e-1
[29 ; 30]	6.447e-3*[age]+-2.105e-1
]30 ; 32]	4.278e-3*[age]+-1.455e-1
[33 ; 35]	3.973e-3*[age]+-1.356e-1
]35 ; 36]	9.495e-3*[age]+-3.289e-1
[37 ; 38]	3.025e-3*[age]+-9.528e-2
]38 ; 39]	3.966e-3*[age]+-1.31e-1
[40 ; 43]	5.014e-4*[age]+5.736e-3
]43 ; 47]	3.872e-4*[age]+1.065e-2
[48 ; 51]	3.872e-4*[age]+1.065e-2

Note

In the case of a continuous variable, the Score Card always includes a number of categories that is higher than in the user defined structure or as given by the parameter *band count* if no user structure has been set. Indeed, the encoding of variables for the Score Card adds target curve points to increase the accuracy of coding according to the training data set. These points split some existing categories and thus increase the number of categories in the Score Card.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Risk Mode

The representation of a model equation is easier to read and to interpret in the **Risk Mode** due to stepwise encoding for ordinal and continuous variables.

In the **Risk Mode**, it is easy to define which category has a negative or positive effect on the risk score and consequently on the odds or on the probability of risk.

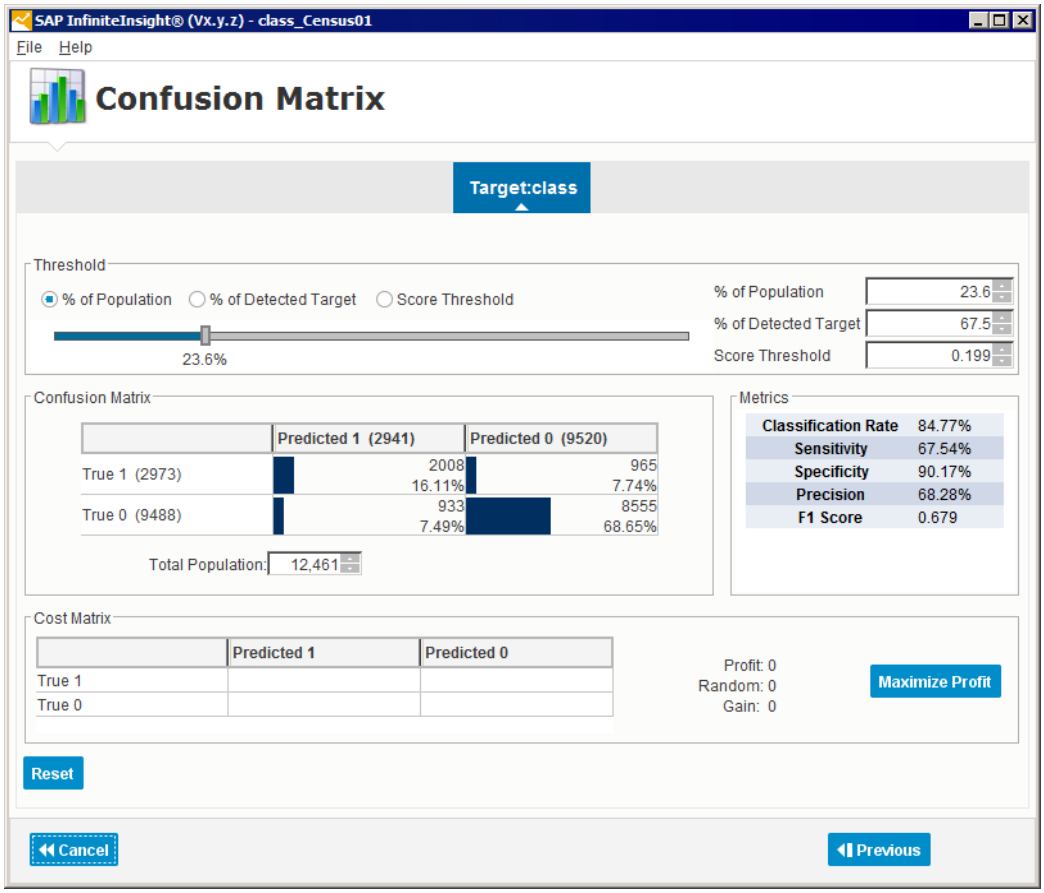
In order to illustrate the advantages of a scorecard in interpreting results, the variable age will be used for this example.

The segment `]24;27]` has a risk score of about 30 and the segment `[37;43]` has a risk score of about 15. According to the parameter PDO (set in this example to 15), it is easy to conclude that the segment `[37;43]` is two times more risky or that the odds of the segment `[37;43]` are two times inferior to the segment `]24;27]`.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Confusion Matrix

The panel *Confusion Matrix* allows you to visualize the target values predicted by the model compared with the real values and to set the score above which the observations will be considered as positive, that is the observations for which the target value is the one wanted. This panel also allows you to simulate your profit depending on the selected threshold score or to automatically adapt the threshold to obtain a maximum profit.



Definitions

- A *positive observation* is an observation that **belongs** to the **target population**.
- A *negative observation* is an observation that **does not belong** to the **target population**.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Understanding the Confusion Matrix

There are three ways to set the threshold using the displayed slide bar:

- by selecting the **percentage of population to target** if the population is sorted by descending order of score (*% of Population*),
- by selecting the **percentage of positive observations** you want to detect (*% of Detected Target*)),
- by selecting the **score** used to differentiate positive observations from negative ones (*Score Threshold*). Any observation with a score above the threshold is considered positive, on the contrary any observation with a score below the threshold is considered negative.

The slide is graduated from the lowest score (on the left) to the highest score (on the right). The values corresponding to each option are displayed is displayed under the slide.

When you move the cursor, the confusion matrix is updated accordingly. The following table details how to read the confusion matrix.

	Predicted[Target Category] Positive Observations Predicted	*Predicted[Non-target Category]* Negative Observations Predicted
True[Target Category] Actual Positive Observations	Number of correctly predicted positive observations	Number of actual positive observations that have been predicted negative
True[Non-target Category] Actual Negative Observations	Number of actual negative observations that have been predicted positive	Number of correctly predicted negative observations

By default, the *Total Population* is the number of records in the Validation data set. You can modify this number to see the confusion matrix for the population on which you want to apply your model.

The Metrics

- The *Classification Rate* is the percentage of data accurately classified by the model when applied on the training data set.
- The *Sensitivity* is the percentage of actual positives which are correctly identified as such.
- The *Specificity* is the percentage of negatives which are correctly identified as such.
- The *Precision* is the percentage to which repeated measurements under unchanged conditions show the same results.
- *The Score* indicates how sensitively a likelihood function depends on its parameter.
  The likelihood of a set of parameter values given some observed outcomes is equal to the probability of those observed outcomes given those parameter values.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Understanding the Cost Matrix

This section allows you to visualize your profit depending on the selected score, or to automatically select the score depending on your profit parameters.

For each observation category, enter a profit or a cost per observation. The total profit is automatically displayed on the right of the table.

To know the threshold that will give you a maximum profit for the profit parameters you have set, click the button *Maximize Profit*.

Example

In the following profit/cost table, each positive observation correctly identified will yield $15, but each negative observation identified as positive will cost you $8.

Category	*Predicted[1]*	*Predicted[0]*
True[1]	15	0
True[0]	-8	0

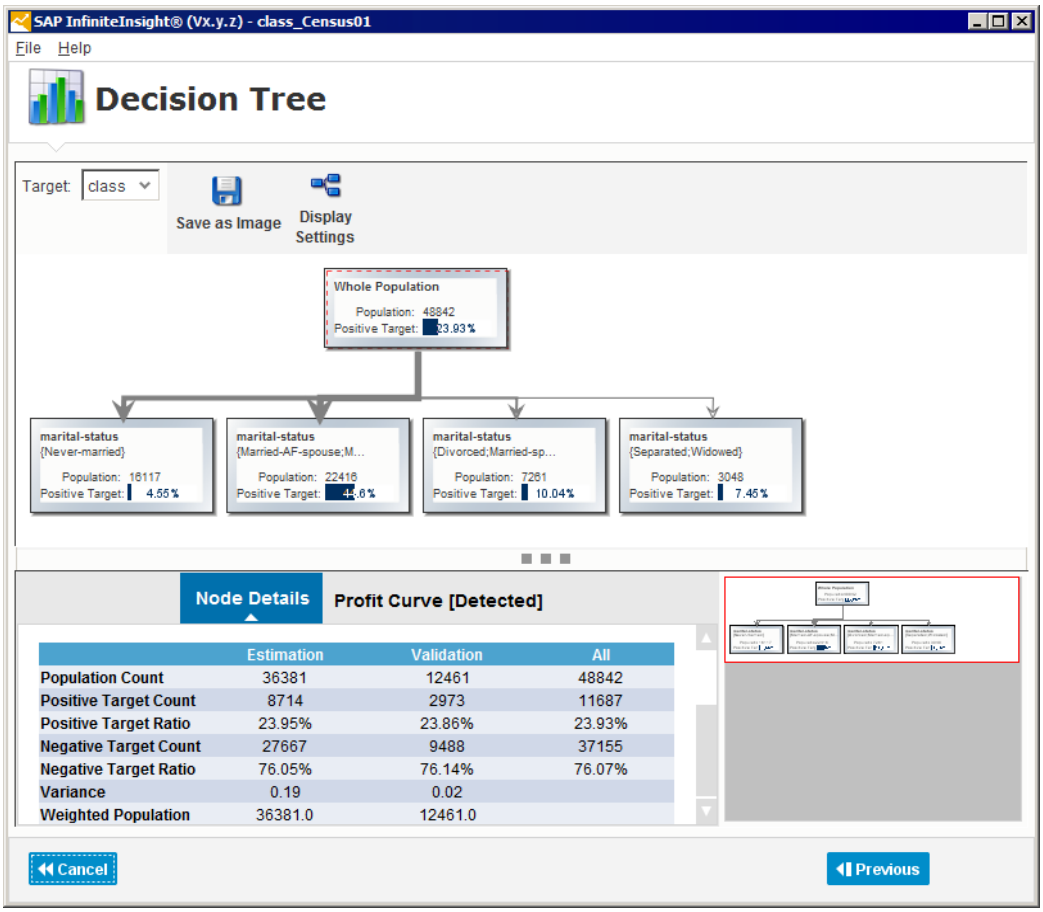
Decision Tree

The panel Decision Tree allows you to display the results of a regression or classification model generated by InfiniteInsight® Modeler as a decision tree based on the five most contributive variables.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Display the Decision Tree for a Target

1 In the *Target* list, select the target for which you want to display the decision tree.

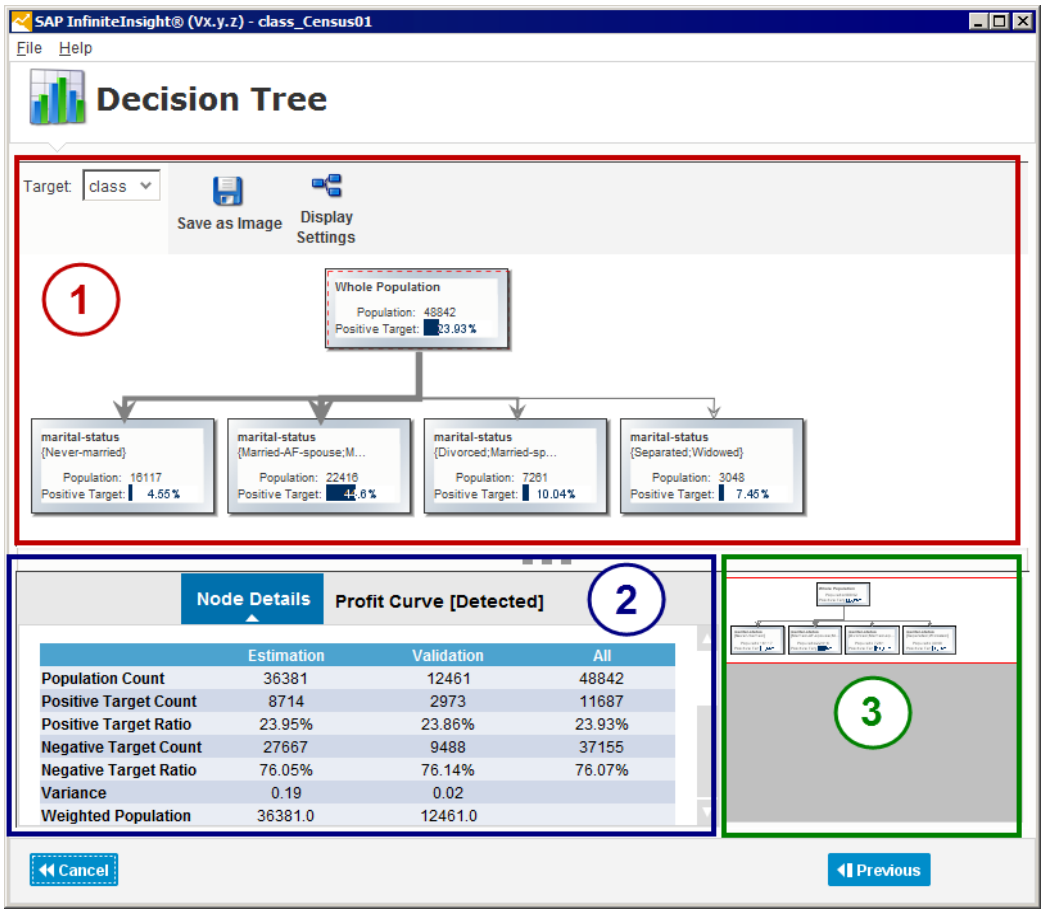


© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Understanding the Decision Tree

The panel *Decision Tree* is split into three parts:

1. the **decision tree** itself, which is displayed in the upper section of the panel,
2. two tabs located in the left bottom part of the panel provide you with **information on the nodes** and with the **profit curve** corresponding to the current decision tree,
3. a **navigator** allowing you to visualize what part of the tree you are studying is displayed in the right bottom part of the panel.

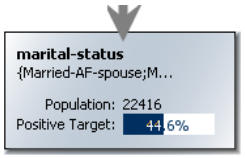
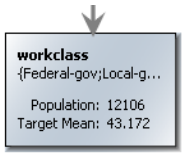


© 2014 SAP AG or an SAP affiliate company. All rights reserved-

The Decision Tree

Each node in the tree displays:

- the name of the expanded variable, for example `Marital-status`.
- the categories on which the node population has been filtered, for example `{Married-AF-spouse;Never-married}`
- the *Population* of the node
- the ratio of *Positive Target* (for nominal targets) or the *Target Mean* (for continuous targets)

Example for a nominal target	*Example for a continuous target*
	

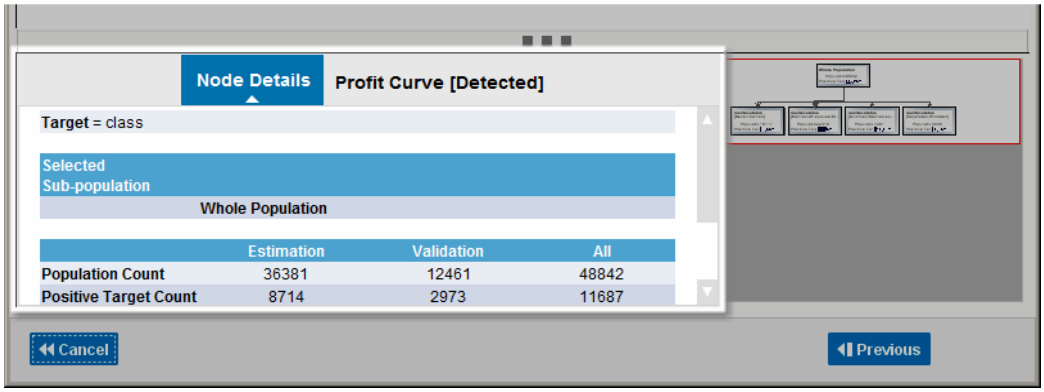
When you go over a node, several options are offered:

	Select the variable to be used to expand the next level of the decision tree.
	Automatically expand the next level using the most contributive variable not yet used in the current decision tree.
	Fold the section of the tree displayed below the current node.

The thickness of the arrows depends on the amount of population in the node. In the following example, the arrow leading to the node corresponding to the category `[0;4386[` of `capital-gain` is thicker since the node population is significantly higher than the one from the node `capital-gain ]4386;41310]`.

Node Details

When you select a node, the node information is displayed in the tab *Node Details* located in the lower part of the panel.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

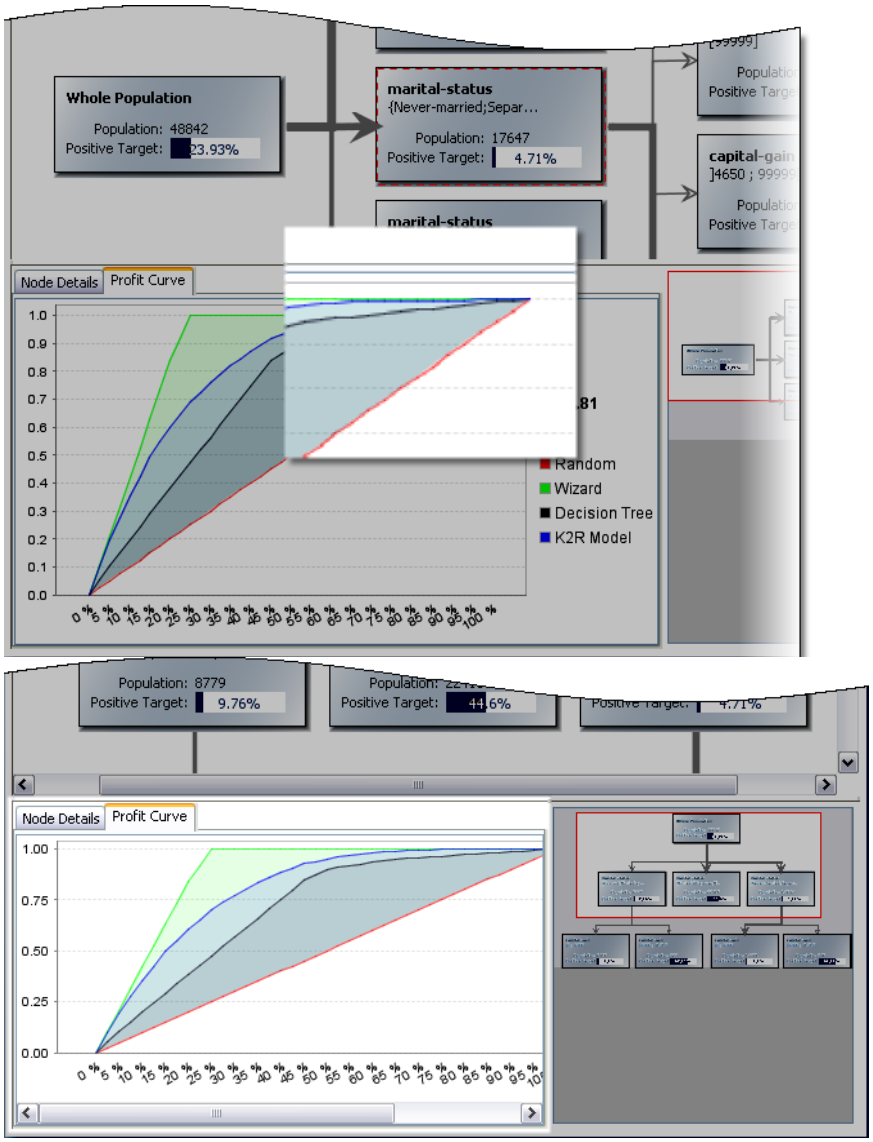
This tab indicates the *target* for which the current decision tree is displayed and provides you with the following information for each data set in the model:

- *Population Count*, that is, the number of records found in the current node
- For continuous targets:
  - *Target Mean*, that is, the mean of the target for the current node.
- For nominal targets:
  - *Positive Target Count*, that is, the number of records for which the target is positive,
  - *Positive Target Ratio*, that is, the percentage of the node population for which the target is positive,
  - *Negative Target Count* , that is, the number of records for which the target is negative,
  - *Negative Target Ratio*, that is, the percentage of the node population for which the target is negative,
  - *Variance*, that is, the variance for the current node,
  - *Weighted Population*, that is, the number of records when using a weight variable.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

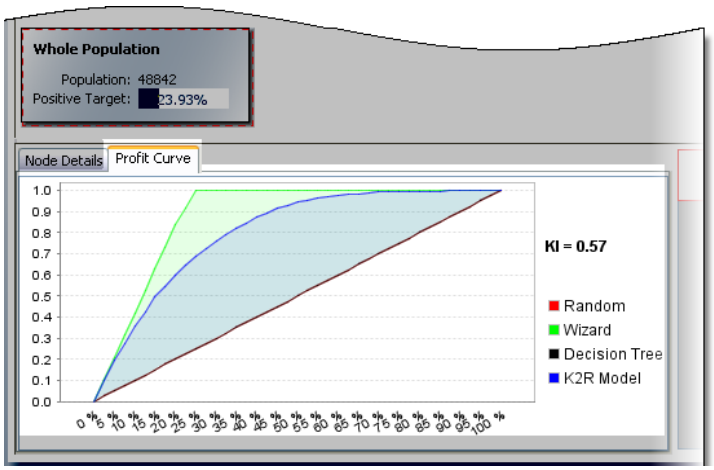
Profit Curve

The profit curve for the current decision tree is displayed in the tab *Profit Curve* located in the lower part of the panel. This profit curve changes with every modifications made on the decision tree.

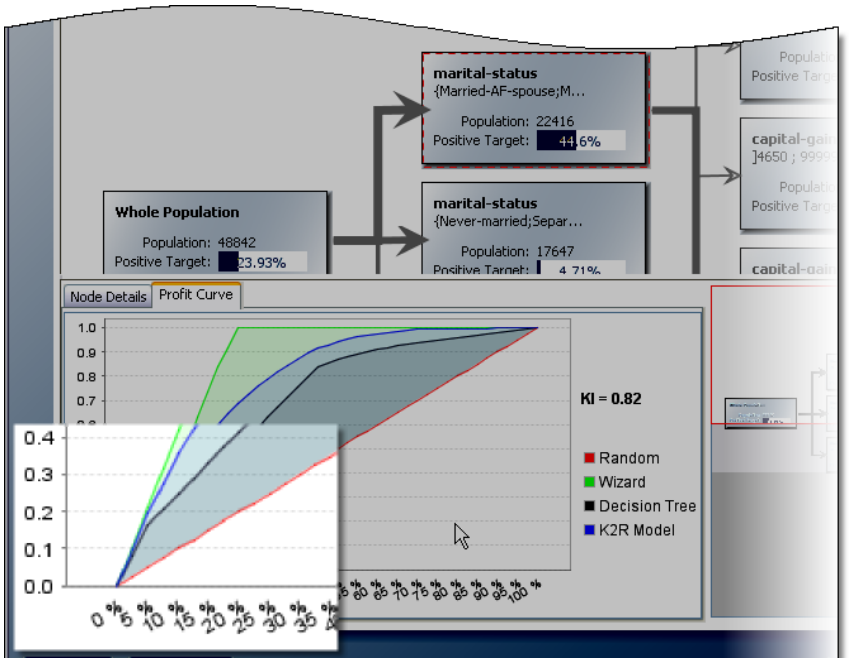


The profit curve corresponding to the node containing the whole population is equal to the random curve.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

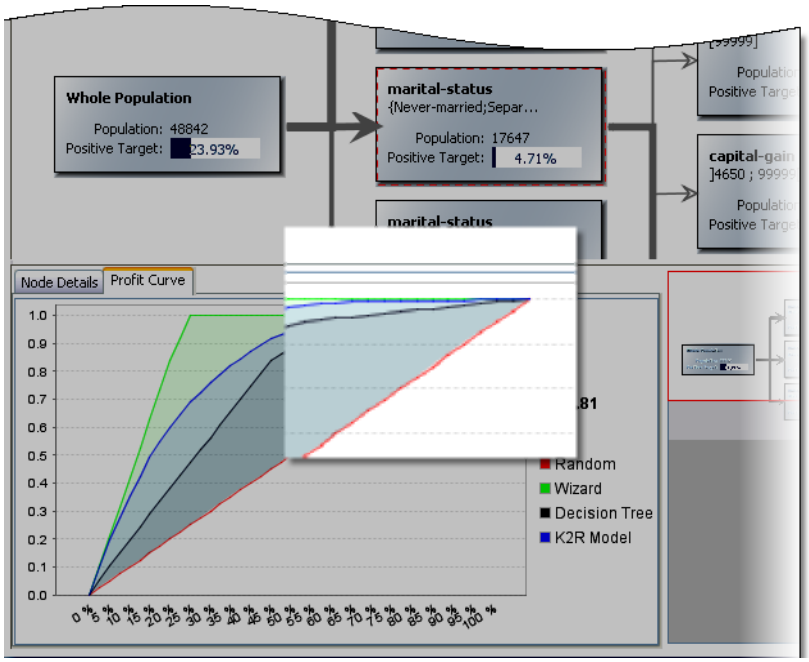


When you expand the node with the **highest percentage of positive target**, the profit curve **improves over the first percentiles**, which means that the model will detect the population with the highest scores.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

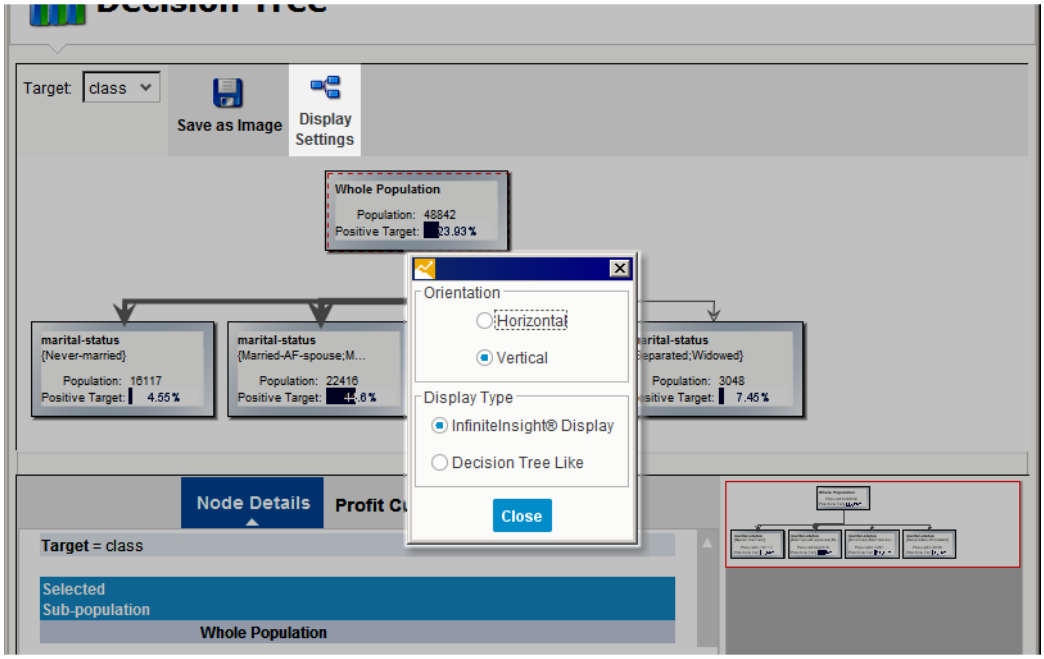
On the contrary, if you expand the node with the **lowest percentage of positive target**, the profit curve will **improve over the last percentiles**.



However, if the node you expand contains a very small population, the profit curve will not be impacted. So you need to find the best compromise between the size of the population and the percentage of positive target.

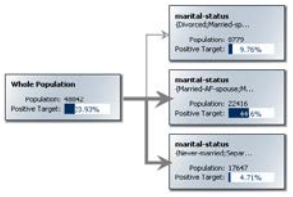
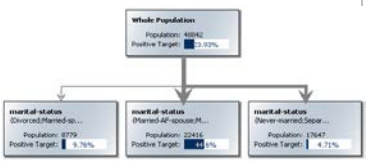
Customizing the Display

The button *Display Settings* allows you to customize some of the display settings for the decision tree.

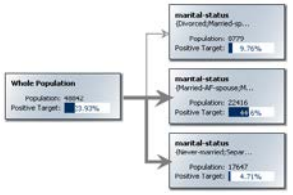
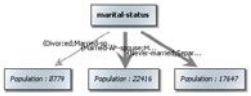


© 2014 SAP AG or an SAP affiliate company. All rights reserved-

- *Orientation*: this setting allows you to select if you want to display the tree horizontally or vertically.

Horizontal	*Vertical*
	

- *Display Type*: this setting allows you to display the decision tree as a standard decision tree (*Decision Tree Like*) or with a specific InfiniteInsight® lo*OK* (*InfiniteInsight® Display*). The option *Decision Tree Like* is more compact but the *InfiniteInsight® Display* is more easily read.

InfiniteInsight® Display	*Decision Tree Like*
	

When you have set the display parameters, click the *Close* button.

5.2.4 Step 4 - Using the Model

Once generated, a classification model may be **saved** for later use.

A classification model may be **applied to additional data sets**. The model thus allows you to perform predictions on these application data sets, by predicting the values of a target variable. The model can also be used to carry out **simulations** on specific observations, on a case-by-case basis.

Moreover, you can **refine** a classification model, by re-generating it with an optimized list of explanatory variables. The SAP InfiniteInsight® allows you to select the variables most pertinent to your business issue automatically, with pertinence defined as: producing the minimum area between the predictive curve and the hypothetical perfect curve, and thus maximizing the volume of information explained by the model.

So that you to apply the model to any other database, SAP InfiniteInsight® allows you to **generate different source code** of the model (**C**, **XML**, **AWK**, **HTML**, **SQL**, **PMML2**, **SAS**, **JAVA, ...**).

Analyzing Deviations

The option *Analyze Deviations* is a tool that provides you with a diagnostic of the data statistical variation.

This option can be used for several purposes:

- to compare the distribution of a new data set with the distribution of the data set used to train the model,
- to check the quality of new data after loading them,
- to check if your data have evolved over time and thus if the model need to be adapted to the new data.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

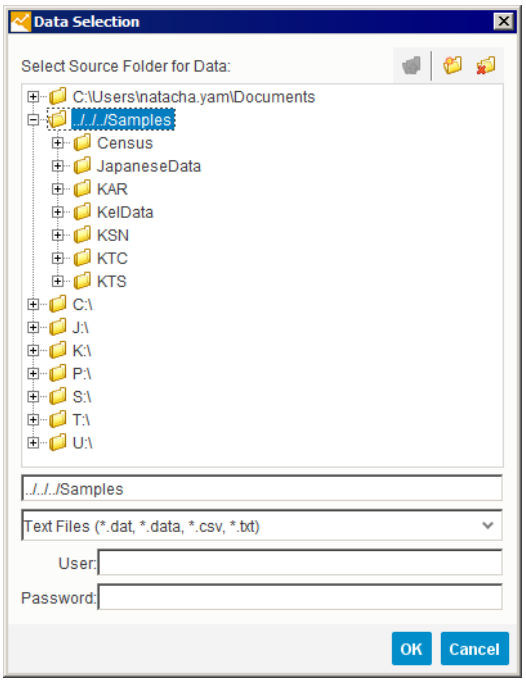
Selecting the Data Set to Analyze

First, you need to select the data set for which you want to analyze the deviations.

For the results to make sense, the new data set should contain the same columns as the data set that was originally used to train the model including the target variable, which must be filled.

☑ To Select a Data Set

1. On the screen *Analyze Deviations*, select the data source format to be used (*Text file, ODBC*, ...).
2. Click the *Browse* button.
   The following selection dialog box appears.

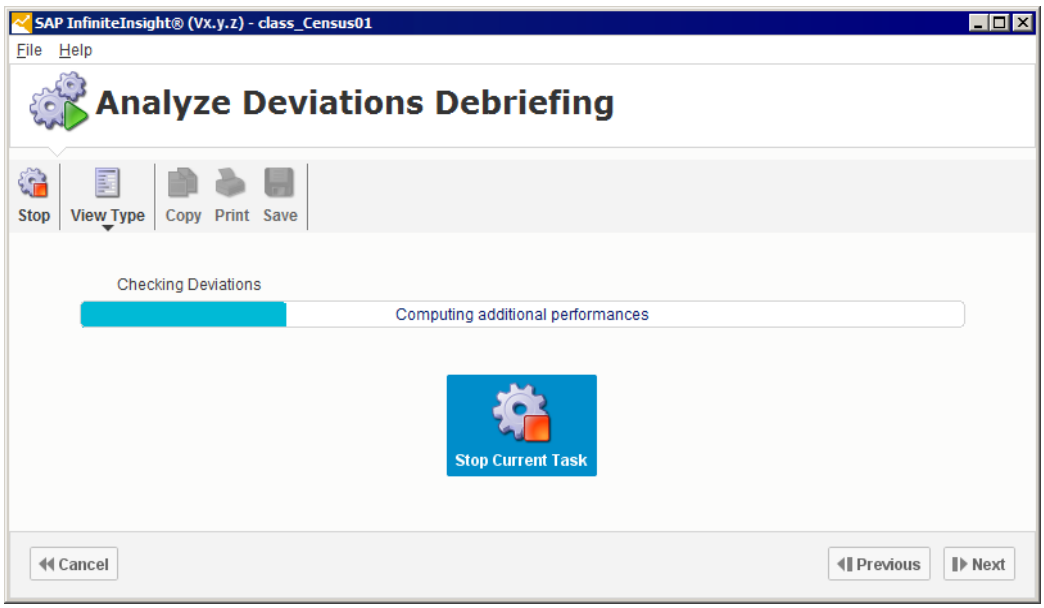


3. Select the file you want to use, then click *OK*.
   The name of the file appears in the *Data Set* field.
4. Click *Next*.
   The screen *Deviation Analysis Debriefing* is displayed.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Following the Deviations Analysis Progress

The panel *Deviation Analysis Debriefing* allows you to follow the analysis process thanks to a progression bar.



At the end of the process, a debriefing panel is displayed. For details on the debriefing panel, see section Understanding the Deviation Analysis (on page 157).

You can use the toolbar provided on the upper part of the panel to:

- stop the analysis process, by clicking the button,
- display the text log detailing the process, by clicking the button,
- copy, print or save the debriefing panel .

☑ To Copy the Report

Click the (*Copy*) button.
The application copies the HTML code of the screen. You can paste into a word processing or spreadsheet program, a text editor, ...

☑ To Print the Report

1. Click the (*Print*) button situated under the title.
   A dialog box appears, allowing you to select the printer to use.
2. Select the printer to use and set other print properties if need be.
3. Click *OK*.
   The report is printed.

☑ To Save the Report

Click the (*Save*) button situated under the title.
The file is saved in HTML format.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Understanding the Deviations Analysis

The first step to take to know if there are any deviations in your data is to lo*OK* at the debriefing report (on page 157) and compare the performances (KI and KR) obtained on the original data with those obtained on the control data set.

Then to visualize which variables have changed, you should lo*OK* into the Control for Deviations Reports (on page 157).

Debriefing Report

The section *Control for Deviation Overview* provides you with basic statistics on the *Data Set used for Deviation Control* (also called **control data set**) such as:

- the name of the data set (*Data Set*),
- the source file (*Source*),
- the number of records contained in the data set (*Number of Records*)
- and the number of variables for which SAP InfiniteInsight® has found deviations in comparison to the data set originally used to train the model (*Number of variables showing deviation*).

The second and third section of the debriefing report allows you to compare the performance of your model on the original data set with the its performance on the control data set:

- the section *Performance Indicators* displays for each target the *KI* and *KR* indicators obtained by the model on the original data set.
- the section *Performance on Control Data Set* displays for each target the *KI* and *KR* indicators obtained by the model on the control data set.

If the KI and/or KR of the model on the control data set are significantly **lower** it means that the relation between the variables and the target variable has changed, as a consequence **the model should be rebuilt** on the new data.

If the KI and KR are not much different, it means that the relation between the input variables and the target behavior has not changed, but it does not mean that differences of distributions are not possible.

Control for Deviations Reports

The panel *Control for Deviations* provides you with six options that can be separated in three groups:

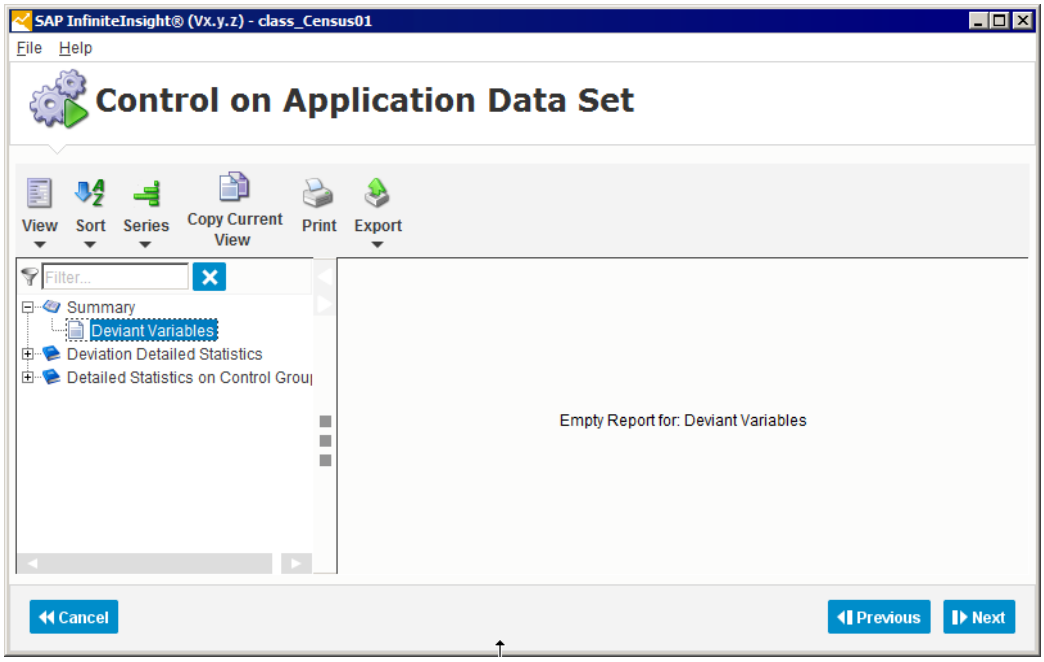
- the first one, made of the options *Probability of Deviation*, *Probability of Category Deviation* and *Probability of Grouped Category Deviation*, enumerates the probabilities of deviation of each variable distribution, be it by variable, variable category or group of categories. A probability over 0.95 indicates that the variable or category global distribution is significantly different in the control data set than in the reference data set.

> **Note**
>
> The probability of deviation is actually a standardized Chi² test. It is significant above 0.95.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

- the second group, comprised of the options *Probability of Target Deviation* and *Probability of Target Deviation for Grouped Categories*, lists for each variable the probabilities of deviation of the categories and the grouped categories with respect to the target variable. A probability over 0.95 indicates that there is a change of behavior with respect to the target variable in the category or group of categories.
- the last group contains only the option *Category with Problem*. For each data set (reference data sets and control data set), all variable categories with a probability over 0.95 are listed. This allows you a quick visualization of possible problems without having to analyze all the reports.

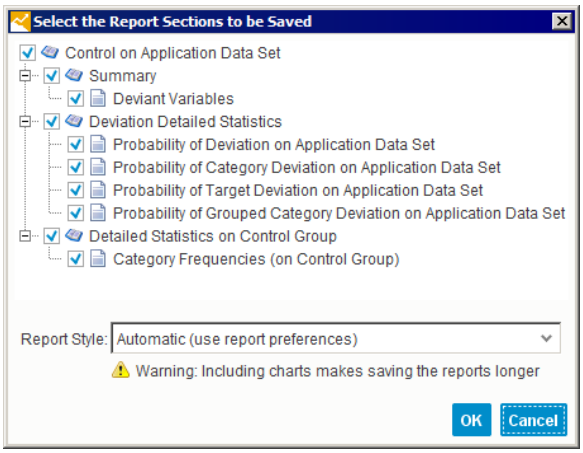
> **Caution**
>
> In all the report panels the control data set is referred as the *ApplyIn* data set.



Options

You can select which report sections to save:

1 Click the button *Save the reports* (located in the bottom left corner). A selector window opens.


Select the Report Sections to be Saved
Control on Application Data Set
Summary
Deviant Variables
Deviation Detailed Statistics
Probability of Deviation on Application Data Set
Probability of Category Deviation on Application Data Set
Probability of Target Deviation on Application Data Set
Probability of Grouped Category Deviation on Application Data Set
Detailed Statistics on Control Group
Category Frequencies (on Control Group)
Report Style:
Automatic (use report preferences)
Warning: Including charts makes saving the reports longer
OK
Cancel

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2. In the displayed list, check the sections you want to save.
3. In the list *Report Style*, select the type of output you want. Three styles of output are available:
   - *Automatic*: saves the default view displayed in the interface
   - *Graphical*: saves the reports as graphs if such a view exists
   - *Textual*: saves the reports as tables
4. Click *OK*.
5. Select the folder in which you want to save the report.
6. Enter the name of the file.

Caution

When selecting the options *Automatic* or *Graphical*, be careful to choose an appropriate file type such as pdf, rtf or HTML.

Applying the Model to a New Data Set

The currently open model may be applied to additional data sets. The model allows you to perform predictions using the application data sets, and specifically, to predict the values of the target variable.

Constraints of Model Use

In order to apply a model to a data set, the format of the application data set must be identical to that of the training data set used to generate the model. The same target variable, in particular, must be included in both data sets, even if the values are not contained in the application data set.

Note

If the *KxIndex* variable of the model is virtual, the application data set must not contain a physical *KxIndex* variable.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

For this Scenario

Due to technical constraints, a data set corresponding to the database of 1,000,000 customers that will be used in this scenario can not be provided to you.

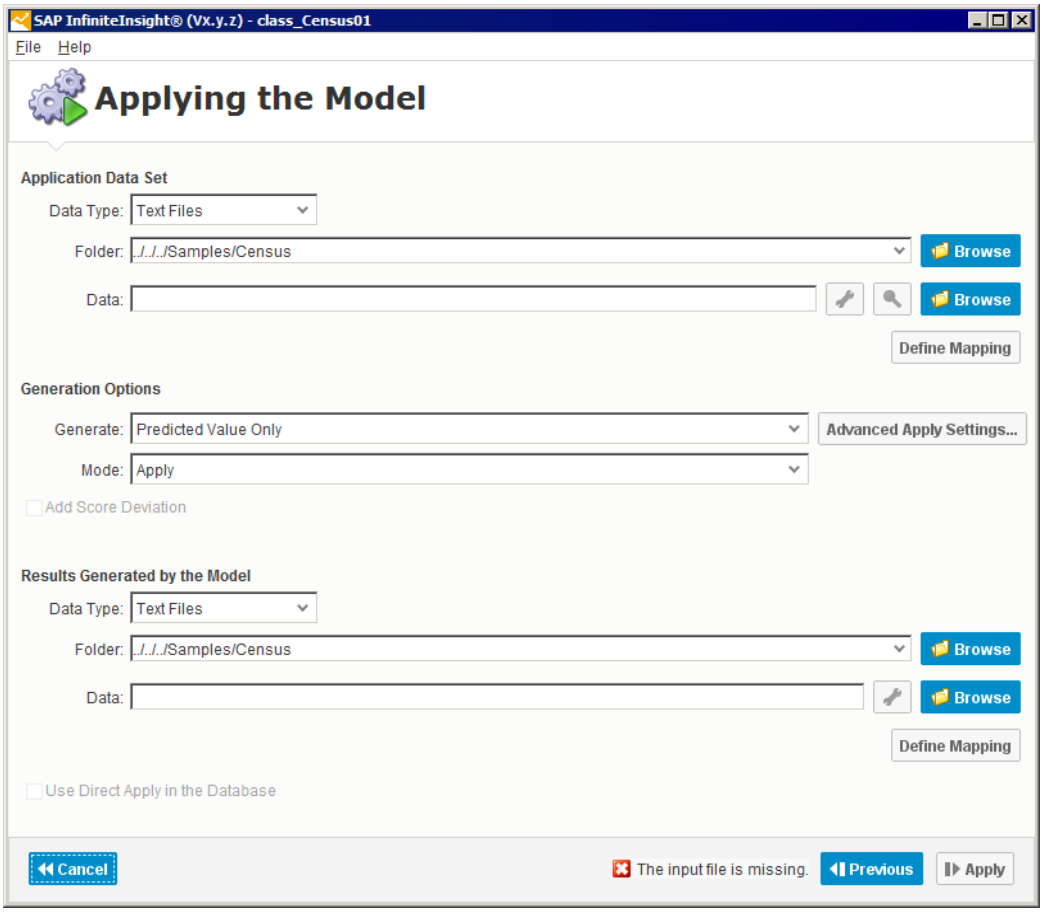
You will apply the model to the file *Census01.csv*, which you used to generate the model. In this manner, you will be able to compare the predictions provided by the model to the real values of the target variable *Class* for each of the observations.

In the procedure *To Apply the Model to a New Data Set*:

- Select the format *Text files*,
- In the *Generate* field, select the option *Individual Contributions*,
- Select the folder of your choice in which to save the results file (*Model-Generated Output*).
- Do not select the option *Keep only outliers*.

To Apply the Model to a New Data Set

1 On the screen *Using the Model*, click the option *Applying the model to a new data set.*
The screen *Applying the Model* appears.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

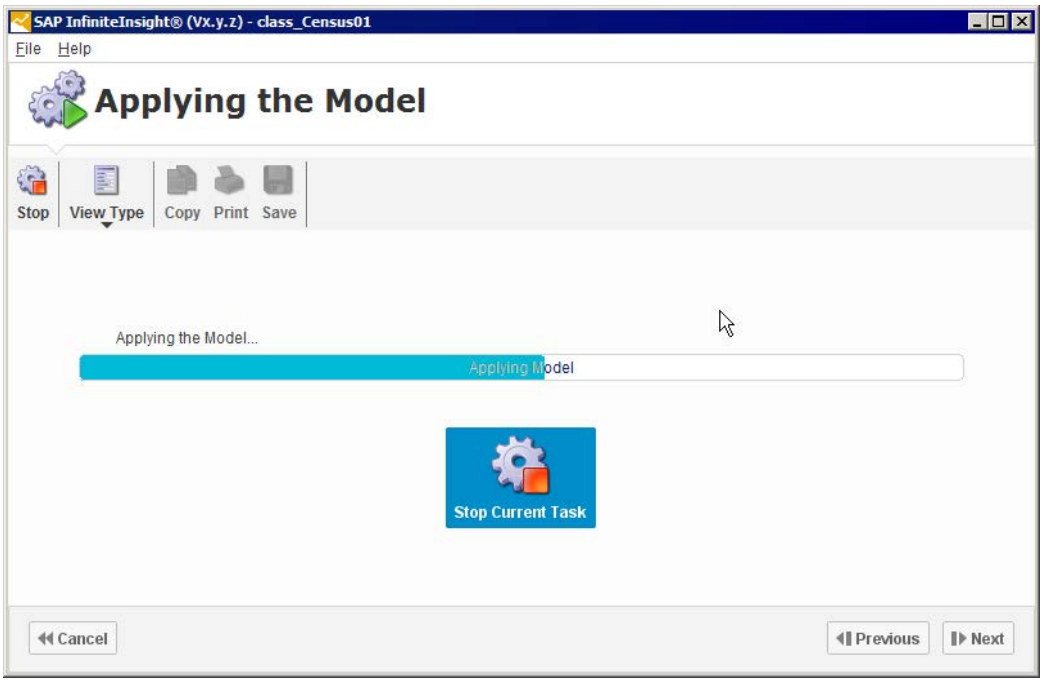
2 In the section *Application Data Set*, select the format of the data source in the list *Data Type*.

3 Click the *Browse* button to select:
- In the *Folder* field, the folder which contains your data set,
- In the *Data* field, the name of the file corresponding to your data set.

4 In the section *Results generated by the model*, select the file format for the output file in the list *Data Type*.

5 You may also opt to select *Keep only outliers*.
If you select this option, only the outlier observations will be presented in the results file obtained, after applying a model.

6 Click the *Apply* button.
The screen *Applying the Model* appears.
Once application of the model has been completed, the results files of the application is automatically saved in the location that you had defined from the screen *Applying the Model*.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Classification Decision

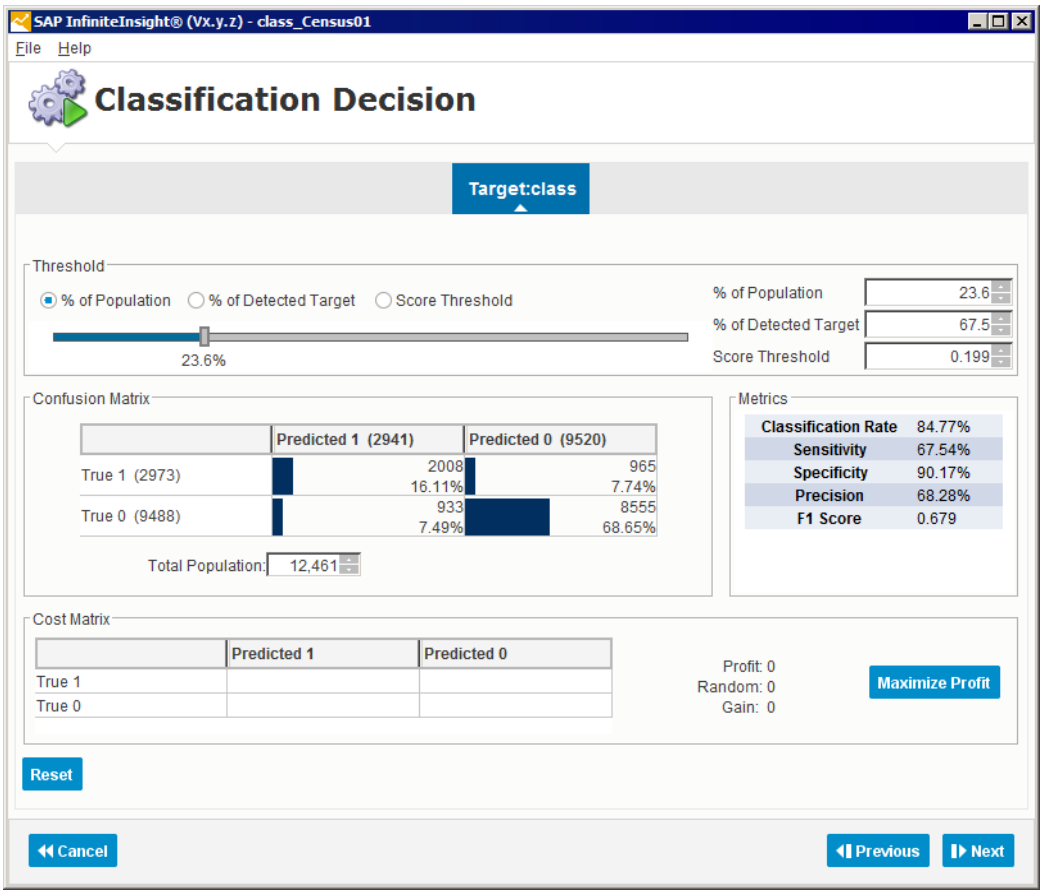
The screen *Classification Decision* allows you to select how many observations you want the model to detect after application on the new data set.

☑ To apply a Classification Decision

1 On the screen *Applying the Model*, follow all the steps of the procedure **To Apply a Model to a New Data Set**.

2 In the *Generate* drop-down list, select the option *Decision*.

3 Click the *Apply* button.
The screen *Classification Decision* appears.



4 Use the slide to set the percentage of population to detect.

5 Click *Next*.
The model is applied to the new data set.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Understanding the Classification Decision Screen

The screen *Classification Decision* allows you to either select a percentage of the population who will respond positively to your campaign (*% of Detected Target*) or a percentage of the entire population (*% of Population*).

When moving the cursor on the scale, the different values are updated accordingly.

For example, if you select the option *% of Detected Target* and set the cursor to 80%, the value of the field *% of Population* will be *32.0*, which means that if you want that 80% of the people who will respond positively to your campaign receive your mailing, you will have to send it to 32% of the entire population.

On the other hand, if you select the option *% of Population* and set the cursor to 20% on the scale, the value of the field *% of Detected Target* will be *60.4*, which means that if your budget only allows you to send your mailing to 20% of the entire population, you will touch 60% of the population who will respond positively.

For more details on how to use the *Confusion Matrix*, see section Analyzing and Understanding the Model > Confusion Matrix (on page 144).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Using the Option Direct Apply in the Database

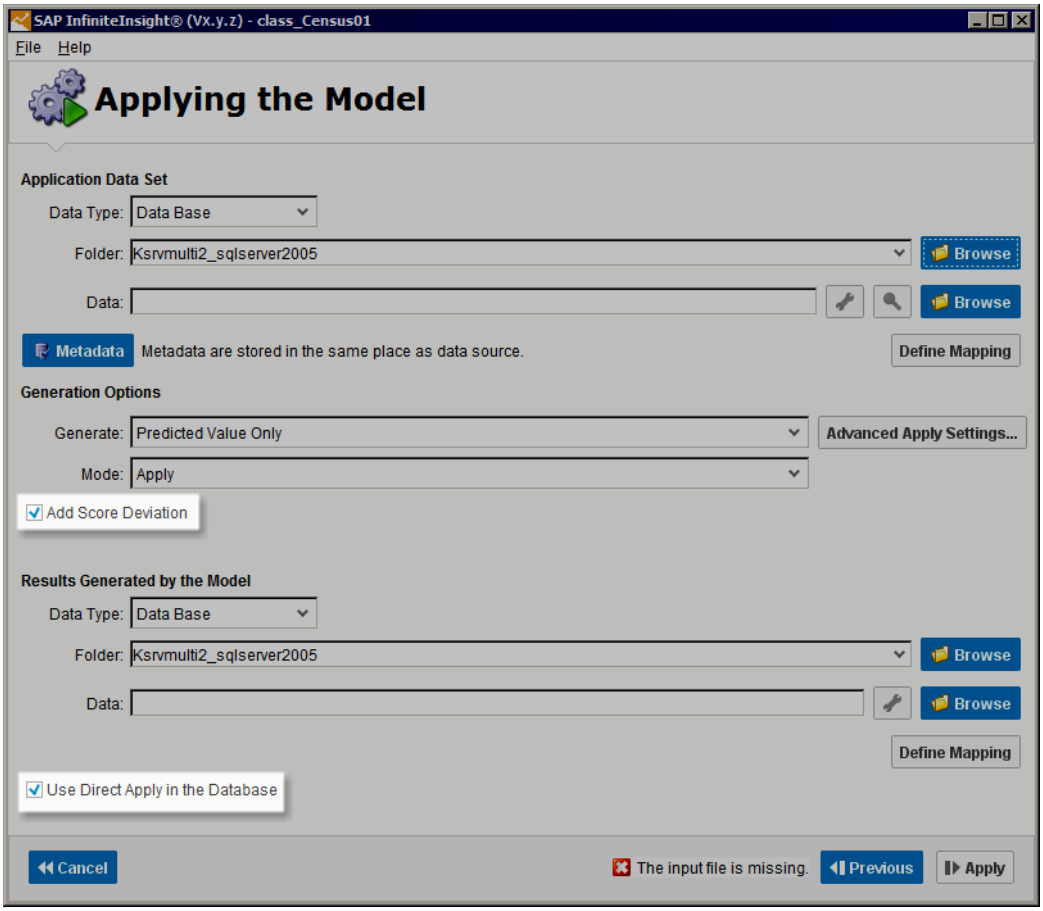
Pre-requisites for the In-database Apply Mode

This optimized scoring mode can be used if all the following conditions are met :

- the apply-in data set (table, view, select statement, data manipulation) and the results data set are tables coming from the same database,
- the model has been computed while at least one physical key variable was defined in SAP InfiniteInsight®,
- there is a valid InfiniteInsight® Scorer license for the database,
- no error has occurred,
- the in-database apply mode is not deactivated,
- granted access to read and write (create table).

☑ To Use the In-database Apply Mode

Check the option **Use the Direct Apply in the Database** and automatically the option **Add Score Deviation** is selected as well.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Advanced Apply Settings

General Outputs

Copy the Weight Variable

This option allows you to add to the output file the weight variable if it had been set during the variable selection of the model.

Copy Data Set Id

This option allows you to add to the output file the name of the sub-data set the record comes from (*Estimation*, *Validation* or *Test*).

Caution

This option cannot be used with the in-database apply feature.

Copy the Variables

This option allows you to add to the output file one or more variables from the data set.

☑ **To Add All the Variables**

Check the *All* option.

☑ **To Select only Specific Variables**

1. Check the *Individual* option.
2. Click the >> button to display the variable selection table.
3. In the *Available* list select the variables you want to add (use the *Ctrl* key to select more than one variable).
4. Click the > button to add the selected variables to the *Selected* list.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

User Defined Constant Outputs

This option allows you to add to the output file constants such as the apply date, the data set name, or any other information useful for using the output file.

A user defined constant is made of the following information:

Parameter	Description	Value / Warnings
Visibility	indicates if the constant will appear in the output or not	*checked*: the constant appears in the output *unchecked*: the constant does not appear in the output
Name	the name of the user defined constant	**1** The name cannot be the same as the name of an existing variable of the reference data set. **2** If the name is the same as an already existing user defined constant, the new constant will **replace** the previous one
Storage	the constant type (number, string, integer, date, ...)	*number* *string* *integer* *date* *datetime*
Value	the value of the constant	date format: `YYYY-MM-DD` datetime format: `YYYY-MM-DD HH:MM:SS`
Key	indicates if the constant is a key variable or identifier for the record. You can declare multiple keys. They will be built according to the indicated order (1-2-3-...).	*0*: the variable is not an identifier *1*: primary identifier *2*: secondary identifier ...

☑ To Define a Constant

1. Click the *Add* button. A pop-up window opens allowing you to set the constant parameters.
2. In the field *Output Name*, enter the constant name.
3. In the list *Output Storage*, select the constant type.
4. In the field *Output Value,* enter the constant value.
5. Click *OK* to create the constant. The new constant appears in the list. You can choose whether to generate the defined constants or not by checking the *Visibility* box.

Gain Chart

This tab allows you to compute the gain chart on the apply data set, that is to rank your data in order of descending scores and split it into exact quantiles (decile, vingtile, percentile).

☑ To Compute the Gain Chart

1. Check the box *Compute Gain Chart on Apply-in Data*.
2. In the list, select the *Number of Quantiles* you want your data to be segmented in.
3. You can add additional variables in order to estimate profits per segments of the population:
   1. In the *Variables* list, select the variables you want to add to the gain chart. Use the *CTRL* key to select multiple variables.
   2. Click the > button to add the selected variables to the list *Values for Gain Chart*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4 The sum of each selected variable will be calculated for each segment of the population.

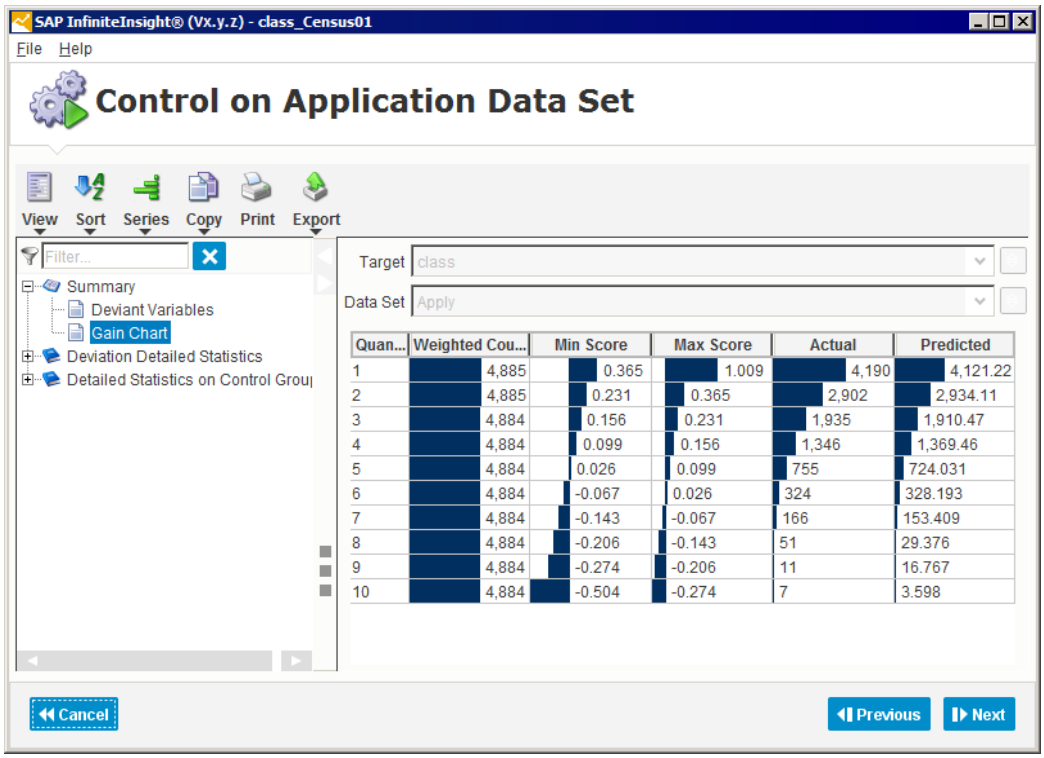
5 Click *Validate* to save the advanced parameters and go back to the panel *Applying the Model*.

☑ Results

The result of the gain chart computation is available at the end of the model application. It can also be found in the *Statistical Reports*, in the section *Model Performance*.

Results

The result of the gain chart computation is available at the end of the model application.


SAP InfiniteInsight® (Vx.y.z) - class_Census01
File Help
Control on Application Data Set
View Sort Series Copy Print Export
Filter...
Summary
Deviant Variables
Gain Chart
Deviation Detailed Statistics
Detailed Statistics on Control Grou
Target class
Data Set Apply
Quan... Weighted Cou... Min Score Max Score Actual Predicted
1 4,885 0.365 1.009 4,190 4,121.22
2 4,885 0.231 0.365 2,902 2,934.11
3 4,884 0.156 0.231 1,935 1,910.47
4 4,884 0.099 0.156 1,346 1,369.46
5 4,884 0.026 0.099 755 724.031
6 4,884 -0.067 0.026 324 328.193
7 4,884 -0.143 -0.067 166 153.409
8 4,884 -0.206 -0.143 51 29.376
9 4,884 -0.274 -0.206 11 16.767
10 4,884 -0.504 -0.274 7 3.598
Cancel
Previous
Next

It can also be found in the *Statistical Reports*, in the section *Model Performance*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Outputs by Targets

Reason Codes

This feature allows retrieving a list of the variables whose values influence the most a score-based decision (typically a risk score). An example of use of the reason codes is to provide a customer with the reasons why the automatic scoring system did not approve their loan.

☑ To Generate Reason Codes

1. In the tree *Advanced Apply Settings* located on the left of the panel, open the node *Outputs for Target '<Target Name>'*.
2. Select *Reason Codes*.
3. Click the + button located on the right of the displayed table.
4. Click in the cell corresponding to the parameter you want to set. The following table sums up the available parameters:

Parameter	Values	Description
Number of Reason Codes	Integer Default: *3*	Number of reason codes you want to generate.
Threshold	▪ *Mean* (default) ▪ Maximum ▪ Minimum	Threshold used for computing the most important reason codes. For each variable the contribution corresponding to the customer score is compared to its contribution for the whole population. The variables for which the contribution is the most differential are selected as the most important reason codes. For example, if you select *Mean*, the customer variable contribution will be compared to the mean of the whole population contribution to determine which variables are the most differential.
Criterion	▪ *Below* (default) ▪ Above	Indicates whether you want to generate the reason codes when the customer variable contribution is above or below the threshold *Warning* - Using *Below* with the *Minimum* threshold or *Above* with the *Maximum* threshold will generate an error.

5. If you want to generate several types of reason codes, repeat steps 3 and 4 for each type.

Output

The output table contains two columns for each reason code requested:

- `reason_name_<criterion>_<threshold>_<rank>_rr_<target name>`: contains the name of the variable selected as a reason code.
  For example, the output column named `reason_name_Below_Mean_1_rr_class` contains the name of the variable being the most important (`1`) reason code with respect to the target variable `class`. Among the variables whose contribution is below (`Below`) the mean (`Mean`) of the population contribution, the selected variable will be the one having the highest deviation with it.
- `reason_value_<criterion>_<threshold>_<rank>_rr_<target name>`: contains the value of the reason code, that is the difference between the variable contribution for the customer and the threshold.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Continuous Target

Option	Output Column Name	This option allows you to...
Predicted Value	*rr_<target variable>*	generate in the output file the value predicted by the model for the target variable. This option is checked by default.
Confidence	*bar_rr_<target variable>*	add to the output file the confidence level for the value that has been predicted, this is also known as the error bar. It is computed with "3" standard deviations on the validation data set and bin per bin. The percentage of population corresponding to the "3" standard deviations is of about 99%. Calculation formula: {TargetMean - 3 (sqrt(TargetVariance)); TargetMean + 3 (sqrt(TargetVariance))} When (sqrt(TargetVariance)) is equal to the Standard Deviation TargetMean +/- Standard Deviation is equal to the Confidence Interval
Outlier Indicator	*outlier_rr_<target variable>*	to show in the output file which observations are outliers. An observation is considered an outlier if the difference between its "predicted value" and its "real value" exceeds the value of the error bar. In other words, the error bar is a deviation measure of the values around the predicted score. Possible values are **1** if the observation is an outlier with respect to the current target, else **0**.
Contributions	*contrib_<variable>_rr_<target variable>* For example, if "marital-status" is an explanatory variable for the target variable "class", the column *contrib_marital-status_rr_class* will be generated in the output file.	add the variables contributions for the current variable to the output file. You can add the contributions of all variables or select only the contributions of specific variables (see procedure below).

☑ To Add All Variables Contributions

Check the *All* option.

☑ To Add Specific Variable Contributions

1. Check the *Individual* option.
2. Click the >> button to display the variable selection table.
3. In the *Available* list, select the variables you want to add (use the *Ctrl* key to select more than one variable).
4. Click the > button to add the selected variables to the *Selected* list.

Predicted Value

This option is checked by default. It allows you to generate in the output file the value predicted by the model for the target variable. It appears in the output file as `rr_<target variable>`.

Outlier Indicator

This option allows you to show in the output file which observations are outliers. An observation is considered an outlier if the difference between its "predicted value" and its "real value" exceeds the value of the error bar. In other words, the error bar is a deviation measure of the values around the predicted score. It appears in the output file as `outlier_rr_<target variable>`. Possible values are **1** if the observation is an outlier with respect to the current target, else **0**.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Predicted Value Quantile

This option allows you to cut the output file in quantiles and to assign to each observation the number of the quantile containing it.

Approximate quantiles are constructed based on the sorted distribution and the boundaries of predicted scores from the validation sample. The score boundaries are used to determine *approximate* quantiles on the apply data set.

> **Note**
>
> - Exact quantile computation would require a full sort of the scores obtained on the apply data set which can be consuming.
>
> - From version V6.0, SAP InfiniteInsight® offers **Gain Chart** option for this purpose.

It appears in the output file as *quantile_rr_<target variable>_<number of quantiles>*, for example for a target variable named "class" and a number of quantiles equal to 10, the generated column will be named *quantile_rr_class_10*.

1. Check the option *Predicted Value Quantiles*.
2. In the field *Number of Quantiles*, enter the number of quantiles you want to create.Check the option *Predicted Value Quantiles*.

Contributions

This option allows you to add the variables contributions for the current variable to the output file. You can add the contributions of all variables or select only the contributions of specific variables.
It appears in the output file as *contrib_<variable>_rr_<target variable>*.
For example, if "marital-status" is an explanatory variable for the target variable "class", the column *contrib_marital-status_rr_class* will be generated in the output file.

☑ To Add All Variables Contributions

Check the *All* option.

☑ To Add Specific Variable Contributions

1. Check the *Individual* option.
2. Click the >> button to display the variable selection table.
3. In the *Available* list select the variables you want to add (use the *Ctrl* key to select more than one variable).
4. Click the > button to add the selected variables to the *Selected* list.

Nominal Target

Outputs by Rank

Scores

This option allows you to generate in the output file the best score(s) for each observation. For each line in the application data set, SAP InfiniteInsight® compares the scores obtained by the current observation for each category of the target variable and displays the best score in the column *best_rr_<Target Variable>_1*, then if several scores have been requested, the second best score is displayed in the column *best_rr_<Target Variable>_2*, the third best in the column *best_rr_<Target Variable>_3*, and so on. When using this option with the Decision option described below, you can link the best score with the category that has obtained it.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Decision

This option allows you to generate in the output file the best decision(s) for each observation. Like for the previous option, the scores obtained for each category of the target variable are compared and the category with the best score for the current record is displayed in the column *decision_rr_<Target Variable>*, if several decisions have been requested, the category with the second best score is displayed in the column *decision_rr_<Target Variable>_2*, the one with the third best score in the column *decision_rr_<Target Variable>_3*, and so on.

Probabilities

This option allows you to generate in the output file the probability of the best decisions for each observation. Like for the previous options, the scores obtained for each category of the target variable are compared and the probability of the category with the best score for the current record is displayed in the column *proba_rr_<Target Variable>*. If several decisions have been requested, the probability of the category with the second best score is displayed in the column *proba_rr_<Target Variable>_2*, the one with the third best score in the column *proba_rr_<Target Variable>_3*, and so on.

Outputs by Reference Category

Score

This option allows you to generate in the output file the score corresponding to each data set line for the different categories of the target variable. You can generate the scores for all the target variable categories or select specific categories.
It appears in the output file as *rr_<Target Variable>* for the target variable key category and *rr_<Target Variable>_<Category>* for its other categories.

☑ To Add the Score of All Target Variable Categories

Check the *All* option.

☑ To Add Only the Scores of Selected Categories

1. Check the *Individual* option.
2. In the *Selection* column, check the boxes corresponding to the categories for which you want to add the score in the output file.

Prediction Probability

This option allows you to generate in the output file the probability for one or more target variable categories, that is for each observation the probability of the target variable value to be the selected category.
It appears in the output file as *proba_rr_<Target Variable>* for the target variable key category and as *proba_rr_<Traget Variable>_<Category>* for the other categories of the target variable.

☑ To Add the Probabilities of All Target Variable Categories

Check the *All* option.

☑ To Add Only the Probabilities of Selected Categories

1. Check the *Individual* option.
2. In the *Selection* column, check the boxes corresponding to the categories for which you want to add the probabilities in the output file.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Miscellaneous Outputs

Outlier Indicator

This option allows you to show in the output file which observations are outliers. An observation is considered an outlier if the difference between its "predicted value" and its "real value" exceeds the value of the error bar. In other words, the error bar is a deviation measure of the values around the predicted score. It appears in the output file as *outlier_rr_<target variable>*. Possible values are **1** if the observation is an outlier with respect to the current target, else **0**.

Predicted Value Quantile

This option allows you to cut the output file in quantiles and to assign to each observation the number of the quantile containing it.

Approximate quantiles are constructed based on the sorted distribution and the boundaries of predicted scores from the validation sample. The score boundaries are used to determine *approximate* quantiles on the apply data set.

> **Note**
>
> - Exact quantile computation would require a full sort of the scores obtained on the apply data set which can be resource consuming.
> - SAP InfiniteInsight® V6.0 offers **Gain Chart** option for this purpose.

It appears in the output file as *quantile_rr_<target variable>_<number of quantiles>*, for example for a target variable named "class" and a number of quantiles equal to 10, the generated column will be named *quantile_rr_class_10*.

1. Check the option *Predicted Value Quantiles*.
2. In the field *Number of Quantiles*, enter the number of quantiles you want to create.

Contributions

This option allows you to add the variables contributions for the current variable to the output file. You can add the contributions of all variables or select only the contributions of specific variables.
It appears in the output file as *contrib_<variable>_rr_<target variable>*.
For example, if "marital-status" is an explanatory variable for the target variable "class", the column *contrib_marital-status_rr_class* will be generated in the output file.

☑ To Add All Variables Contributions

Check the *All* option.

☑ To Add Specific Variable Contributions

1. Check the *Individual* option.
2. Click the >> button to display the variable selection table.
3. In the *Available* list select the variables you want to add (use the *Ctrl* key to select more than one variable).
4. Click the > button to add the selected variables to the *Selected* list.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Types of Results Available

The application of a model to a data set allows you to obtain four types of results, which are described in the following table.

Type of Results	Description
score, or predicted value	For a continuous variable, the predicted value corresponds to the value predicted by the model for the target variable of each observation. The "predicted values" correspond to the values read off the X axis of the profit curve plot. The "predicted value" of each observation is calculated by replacing the parameters of the polynomial representing the model, by the values of each of the variables of that observation. For a binary variable, the model outputs a score.
probability	Corresponds to the **probability** of each observation belonging or not to the target category of the target variable.
prediction range, or maximum error	The prediction range allows you to identify outlier observations. An observation is considered an outlier if the difference between its "predicted value" and its "real value" exceeds the value of the prediction range. In other words, the prediction range is a deviation measure of the values aroud the predicted score.
individual contributions	The **individual contributions** by variables contained in the data set with respect to the target variable. The sum of all those individual contributions corresponds with the predicted value (score) to the nearest whole number.
decision	The "decision" option can only be used for classification models, that is, when the target variable is nominal. It allows to generate a classification decision based on the "scores" (or "predicted values") generated by the model. The result file obtained contains a column in which a category of the target variable is assigned to every observation. The decision is taken on the basis of a threshold that is applied on the scores generated by the model. The target category of the target variable is assigned to observations whose scores are superior to the threshold. The default threshold (computed during the generation " or training " of the model) is chosen so that the way the categories of the target variable are assigned to observations is representative from their distribution in the training data set.

Upon the level of information desired, you can choose to generate among several results' files, described in the table below.

Selecting the option...	Will generate a results' file containing the following information...
Predicted value only	Only the predicted value of observations (*rr_TargetVariableName*)
Probability	▪ the **predicted value** ▪ the **probability** (*proba_rr_TargetVariableName*) ▪ the **prediction range** (*bar_rr_TargetVariableName*)
Individual Contributions	▪ the **predicted value** ▪ the **probability** ▪ the **prediction range** ▪ the **individual contributions** of variables (*contrib_VariableName_rr_TargetVariableName*)
Decision	▪ the **predicted value** ▪ the **decision** (*decision_rr_TargetVariableName*) ▪ the **decision probability** (*proba_decision_rr_TargetVariableName*) ▪ the **probability**

For this Scenario

Due to technical constraints, a data set corresponding to the database of 1,000,000 customers that will be used in this scenario cannot be provided to you.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

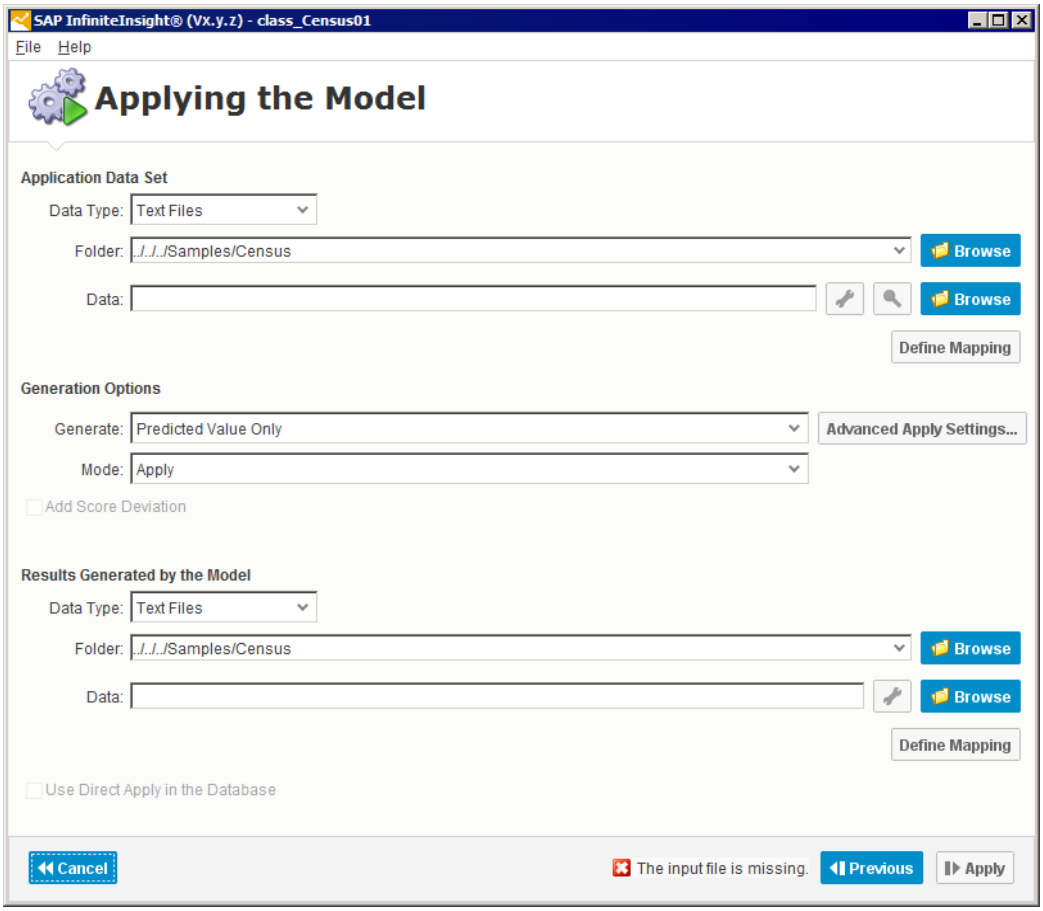
You will apply the model to the file *Census01.csv*, which you used to generate the model. In this manner, you will be able to compare the predictions provided by the model to the real values of the target variable Class for each of the observations.

In the procedure *To Apply the Model to a New Data Set*:

- Select the format *Text files*,
- In the *Generate* field, select the option *Individual Contributions*,
- Select the folder of your choice in which to save the results file (*Model-Generated Output*).
- Do not select the option *Save only outlier observations*.

☑ To Apply the Model to a New Data Set

1 On the screen *Using the Model*, click the option *Apply model*.
The screen *Applying the Model* appears.



2 In the section *Application Data Set*, select the format of the data source.

3 Click the *Browse* button to select:

- In the *Folder* field, the folder which contains your data set,
- In the *Data* field, the name of the file corresponding to your data set.

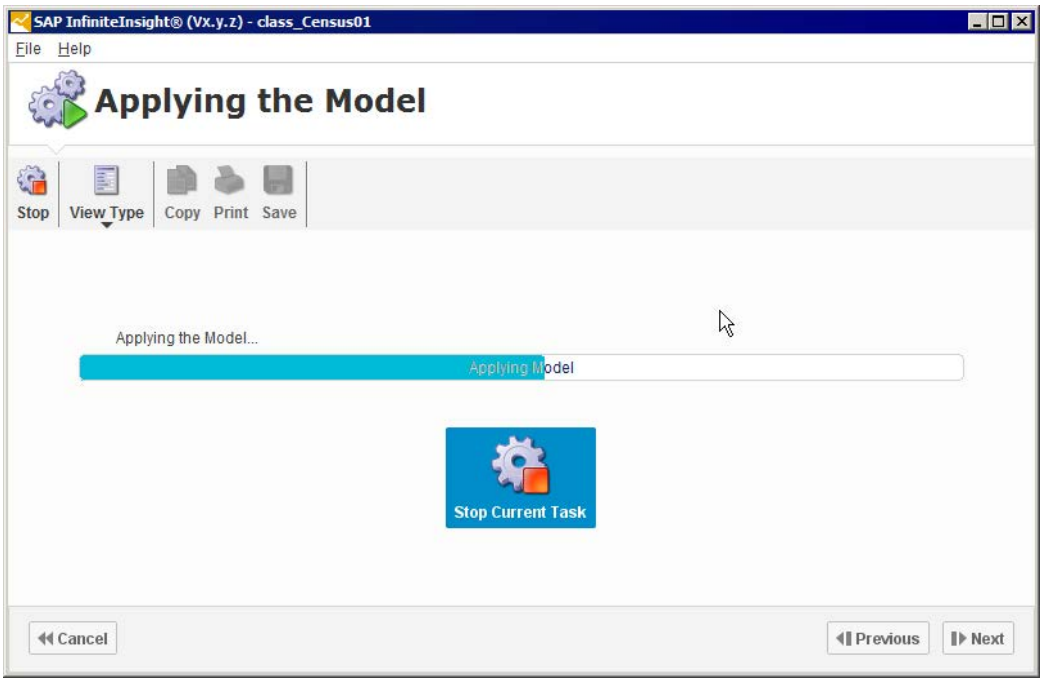
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4 In the section *Results generated by the model*, select the file format for the output file.

> **Note**
>
> The current version of SAP InfiniteInsight® does not allow you to save the file in an ODBC database.
> Click the *Browse* button to select:
> In the *Folder* field, the folder in which you want to save the output file,
> In the *Data* field, the name that you want to give the output file. In the *Generate* Field, select the type of output values that you want for the target variable.

5 You may also opt to select Save only outlier observations.
If you select this option, only the outlier observations will be presented in the results file obtained, after applying a model.

6 Click the *Apply* button.
The screen *Applying the Model* appears.
Once application of the model has been completed, the results files of the application is automatically saved in the location that you had defined from the screen *Applying the Model*.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

For this Scenario

Open the results file in Microsoft Excel, in the text format that you obtained when you applied the model to the *Census01.csv* file.

To Open the Model Application Results File

1 Depending upon the format of the results file generated, use *Microsoft Excel* or another application to open the file.
The figure below presents the headings and columns of the results file obtained for this scenario.

	A	B	C	D	E	F	G
1	KxIndex	class	rr_class	proba_rr_class	bar_rr_class	contrib_age	contrib_workclass
2	1	0	0,02628554	0,177313641	1,17675817	-0,00014918	0,004575992
3	2	0	0,268730283	0,567965508	1,496288657	-0,00604592	0,004575992
4	3	0	-0,192986876	0,023777327	0,457313478	0,00038689	-0,002218456
5	4	0	-0,027590154	0,094321787	0,946476877	-0,00765412	-0,002218456
6	5	0	0,230136439	0,548656046	1,521923184	0,00574755	-0,002218456
7	6	0	0,35434714	0,737949133	1,401365995	0,00092295	-0,002218456
8	7	0	-0,358695984	0,00095057	0,467965275	-0,00550985	-0,002218456
9	8	1	0,188632727	0,435952127	1,483268023	-0,00711805	0,004575992
10	9	1	0,634126425	0,962523282	0,575068474	0,00413935	-0,002218456
11	10	1	0,540619075	0,916772068	0,933868706	-0,00175738	-0,002218456
12	11	1	0,235637605	0,551528811	1,519493461	0,00092295	-0,002218456
13	12	1	0,317775369	0,676707566	1,426785707	0,00467542	0,004575992
14	13	0	-0,192047745	0,024259994	0,463258654	0,00842789	-0,002218456
15	14	0	-0,11635045	0,058997665	0,731234848	0,00360329	-0,002218456
16	15	1	0,164729878	0,406741321	1,466872215	-0,00068525	-0,002218456

2 You can now analyze the results obtained and use these results of your analysis to make the right decisions.

Description of the Results File

Depending upon which options you selected, the results file will contain some or all of the following information, in the same order as seen below:

- The **key variable** defined during data description at the setting model parameters step.
- Possibly, the **target variable** given as known values if the latter appeared in the application data set, as is the case in this scenario.
- The **predicted value** (score) provided by the model for the target variable of each observation. The name of this column corresponds to the name of the target variable prefixed by *rr_*, or in this case, *rr_Class*.
- The **decision** is based on the score. For example, its value can be of 1 if the observation is considered as interesting or 0 if it is considered as uninteresting for the model. The name of this column corresponds to the name of the target variable prefixed by *decision_rr_*, or in this case, *decision_rr_class*.
- The **probability decision** is also based on the score and provides the probability of the decision. The higher it is, the more it will confirm the decision value. The name of this column corresponds to the name of the target variable prefixed by *proba_decision_rr_*, or in this case, *proba_decision_rr_class*.
- The **probability** for each observation that it does or does not belong to the target category of the target variable. The name of this column corresponds to the name of the target variable prefixed by *proba_rr_*, or in this case, *proba_rr_class*.
- The **prediction range**, or "maximum error". The name of this column corresponds to the name of the target variable prefixed by *bar_rr_*, or in this case, *bar_rr_Class*.
- The **individual contributions** by variables contained in the data set with respect to the target variable. The names of the columns of individual contributions correspond to the names of each of the variables, prefixed by *contrib_*, or in this case, *contrib_age*, *contrib_workclass*, and so on.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Performing a Simulation

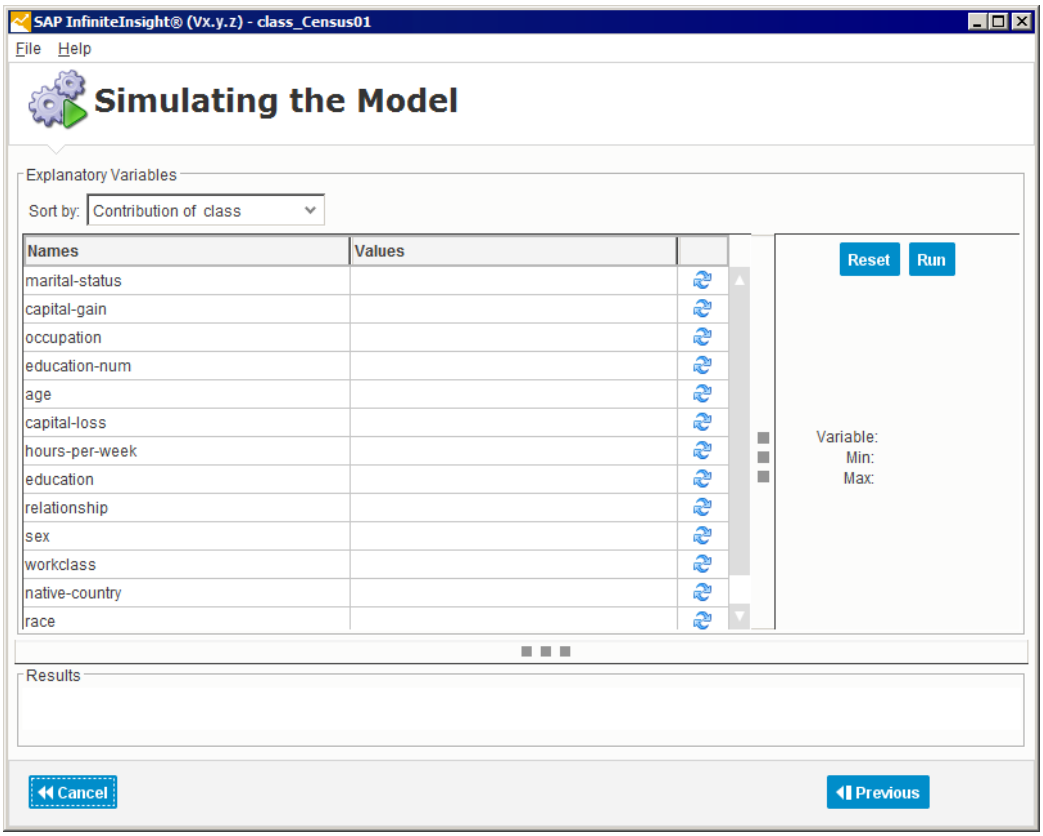
The open model may be used to carry out simulations on specific observations, one at a time. To define the observation to be analyzed, the variables of your choice must be associated with values. For instance, if you have selected the *occupation* (profession category) and *workclass* (socio-professional category) variables, they must contain values. During execution of the simulation, SAP InfiniteInsight® will automatically assign values to certain variables when values are missing, but essential to proper completion of the simulation.

Once the simulation is complete, you will obtain the following results:

- The **predicted value** (score),
- The **probability** that this observation belongs to the target category of the target variable.

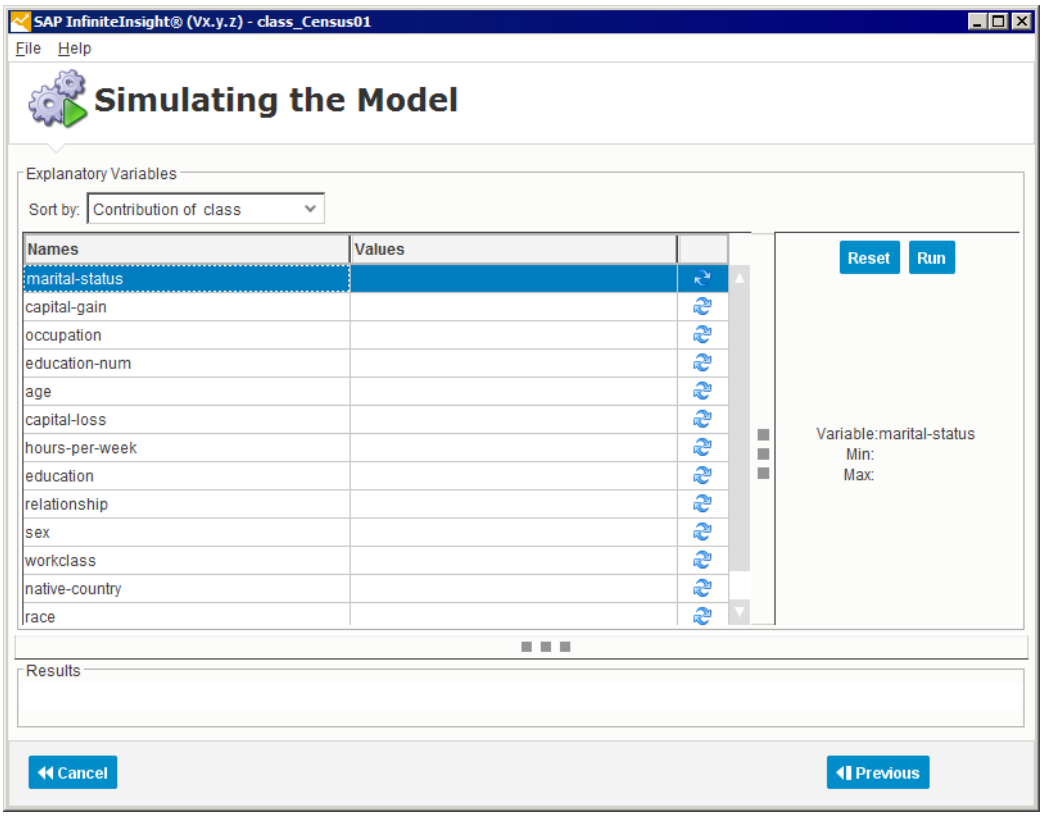
☑ To Simulate a Model

1 On the screen *Using the Model*, click the option *Simulation*.
The screen *Simulating the Model* appears.



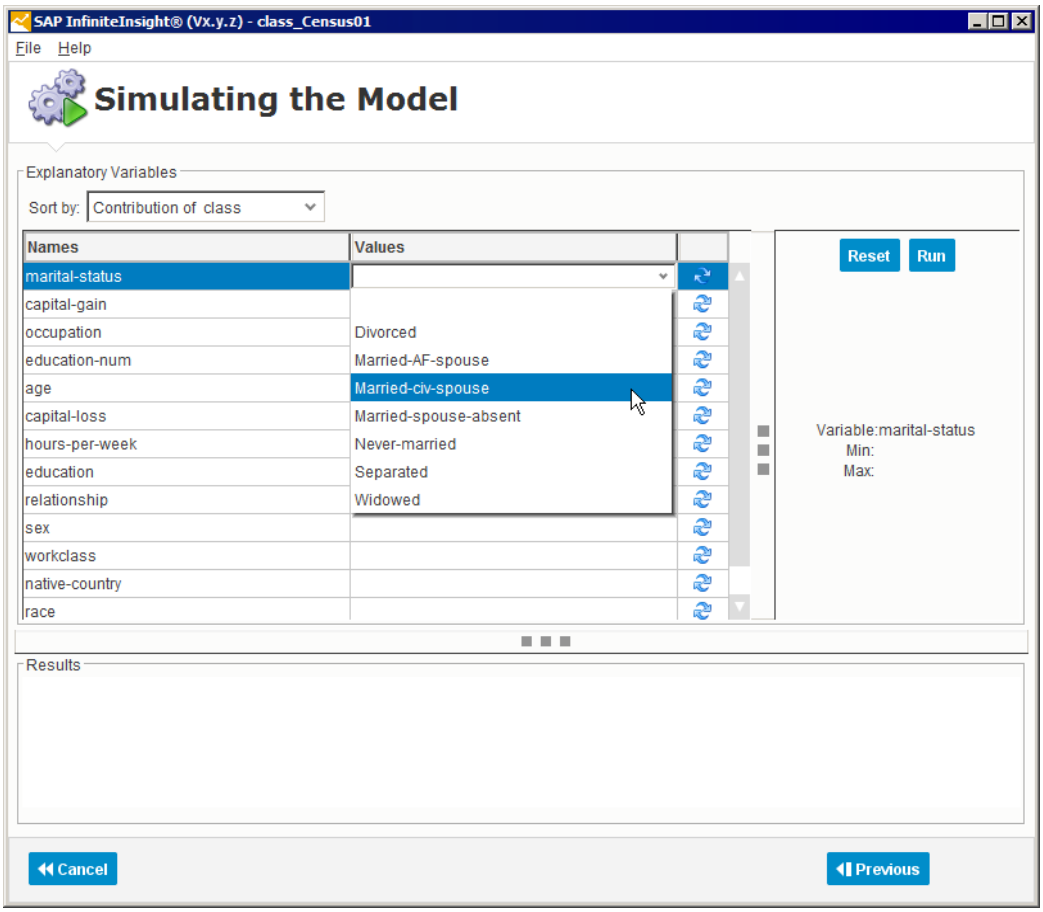
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2 On the left side of the screen (*Explanatory variables*), select a variable, such as *marital-status*. Its values appears in the section *Modifying values*, on the right side of the screen.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

3 In the section *Modifying values*, in the *Value* field, select or enter a value, such as *Married-civ-spouse.* The value appears in the table of *Explanatory variables*, across from the selected variable.



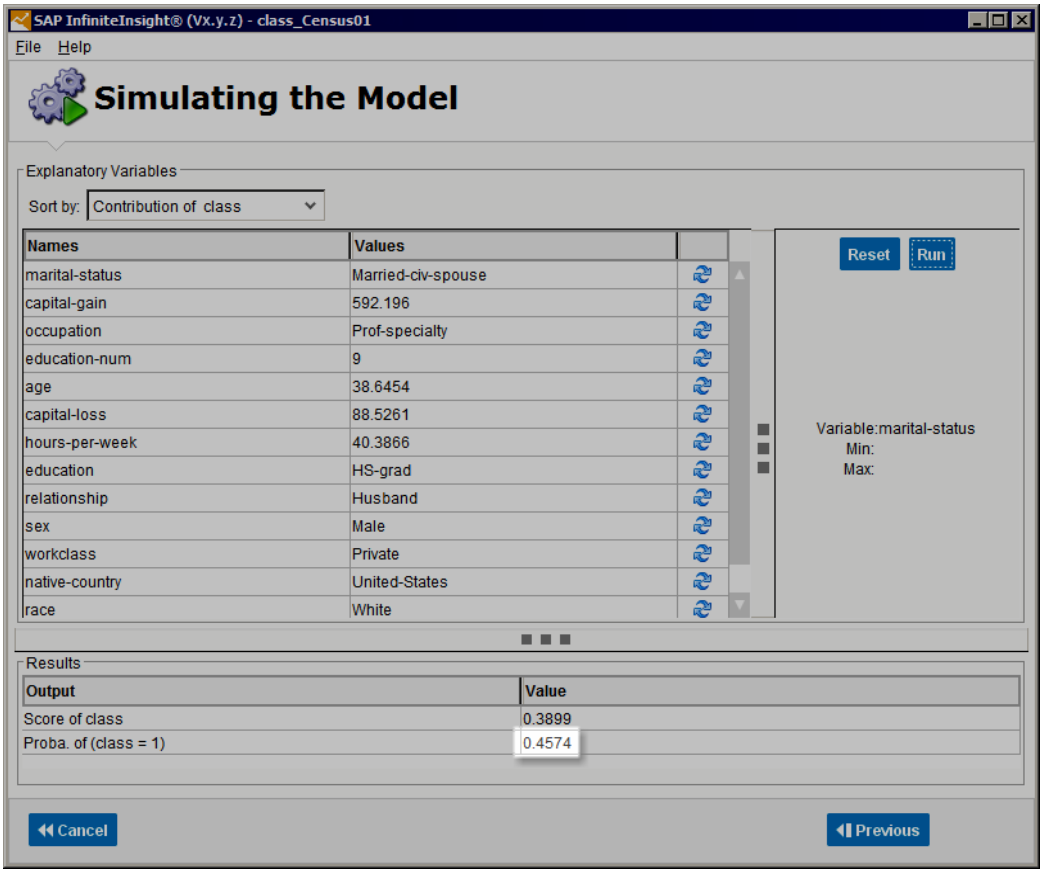
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4 If you would like to select other explanatory variables, go back to step *2*.
Otherwise, go to step *5*.

5 Click the *Run* button to perform a model simulation.
The results of the simulation appears in the *Results* section. You will obtain the *Predicted value* (score) of the observation described in the table of *Explanatory variables*, as well as the *probability* that this observation belongs to the target category of the target variable. In our example, only one variable has been defined. The probability that this observation belongs to the target category of the target variable is 0.1120.
Note that certain variables of the table of *Explanatory variables* were automatically completed upon execution of the simulation. In fact, the model automatically completed certain missing values that were essential to the simulation.
These values are listing in the following table.

Type of variable	Default value
continuous variable	the mean value
nominal variable	the most frequent category
ordinal variable	the most frequent category

These changes are reflected in the left part of the screen after clicking the *Run* button.



6 You can modify the value of an explanatory variable and run the simulation again to measure the effect of that change with respect to the target variable. For instance:

1. Assign the value *Widowed* to the variable *marital-status* in place of the value *Married-civ-spouse*.
2. Run the simulation.
The probability now obtained is 0.0040.

7 Click the *Reset* button to run the simulation again.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Refining a Model

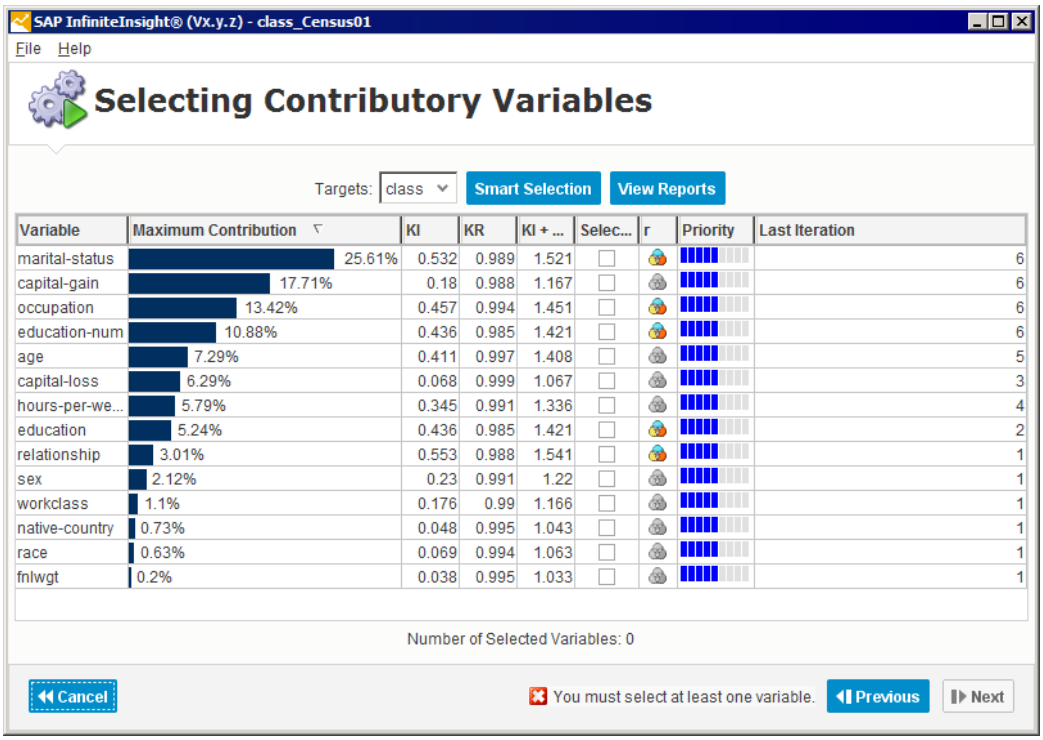
SAP InfiniteInsight® allows you to refine a currently open model. For instance, you can:

- **Reduce the number of explanatory variables** used by the model while maintaining the initial quality (KI) and robustness (KR),
- **Generate a model of degree 2** using the most significant variables of the degree 1 model.

> **Note**
>
> If your data set contains date or datetime variables, automatically generated variables will appear in this panel. For more information, refer to section Date and Datetime Variables: Automatically Generated Variables (on page 30).

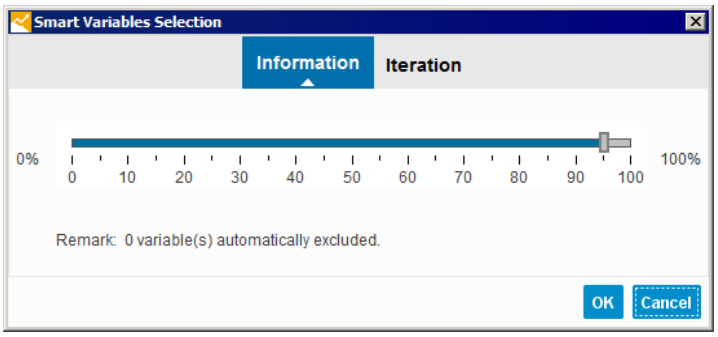
☑ To Refine a Model

1 On the screen *Using the Model*, click the option *Select Variables*.
The screen *Selecting Contributory Variables* appears.


SAP InfiniteInsight® (Vx.y.z) - class_Census01
File Help
Selecting Contributory Variables
Targets: class
Smart Selection
View Reports
Variable | Maximum Contribution | KI | KR | KI + ... | Selec... | r | Priority | Last Iteration
marital-status 25.61% 0.532 0.989 1.521 6
capital-gain 17.71% 0.18 0.988 1.167 6
occupation 13.42% 0.457 0.994 1.451 6
education-num 10.88% 0.436 0.985 1.421 6
age 7.29% 0.411 0.997 1.408 5
capital-loss 6.29% 0.068 0.999 1.067 3
hours-per-we... 5.79% 0.345 0.991 1.336 4
education 5.24% 0.436 0.985 1.421 2
relationship 3.01% 0.553 0.988 1.541 1
sex 2.12% 0.23 0.991 1.22 1
workclass 1.1% 0.176 0.99 1.166 1
native-country 0.73% 0.048 0.995 1.043 1
race 0.63% 0.069 0.994 1.063 1
fnlwgt 0.2% 0.038 0.995 1.033 1
Number of Selected Variables: 0
Cancel
You must select at least one variable.
Previous
Next

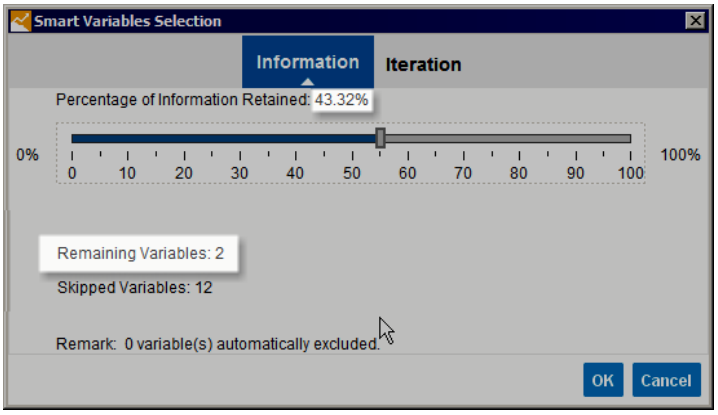
2 In the *Targets* list, select the target variable for which you want to select the contributory variables.

3 Click the button *Smart Selection*. The window *Smart Variables Selection* opens.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4 On the bar *Percentage of Information Retained*, move the cursor to change the amount of information to keep; the number of variables selected changes accordingly.
The further this cursor is moved to the left, the more variables are excluded. The variables excluded are selected automatically as a function of their significance with respect to the model.
For instance, the figure below shows that to retain only two variables out of the original fourteen, you should keep 43.07% of the information contributed by the model.

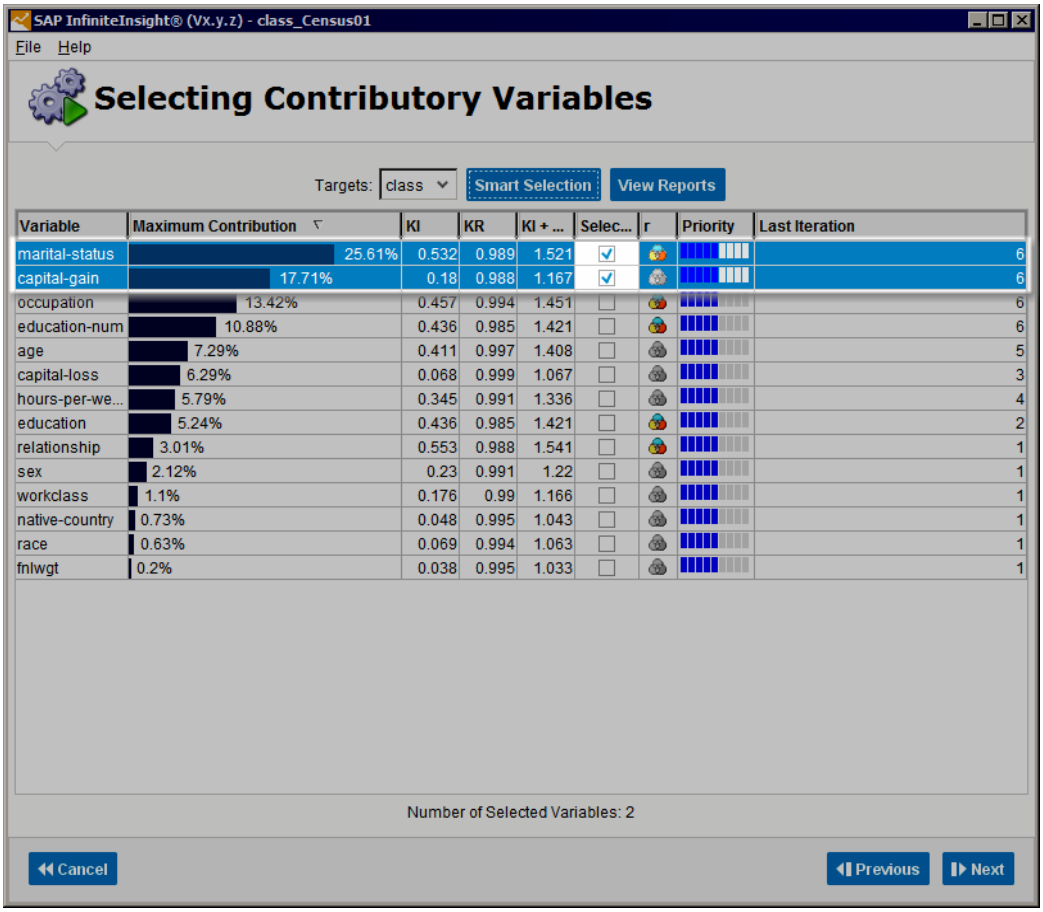


Note

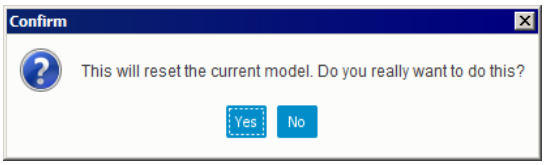
Certain variables in the training data set may contribute no information, such as constant value variables. These can therefore be automatically excluded from the model during the training phase. The number of variables excluded is displayed as a *Remark*. In the figure above, this number is equal to "0".

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

5 Click *OK*.
The window *Smart Variables Selection* closes and the panel *Selecting Contributory Variables* is updated with the selected variables, allowing you to view the kept variables and the excluded ones. In our example, SAP InfiniteInsight® automatically determined that the two explanatory variables that contributed the most information to explain the target variable were the variables *marital-status*, and *capital-gain*.

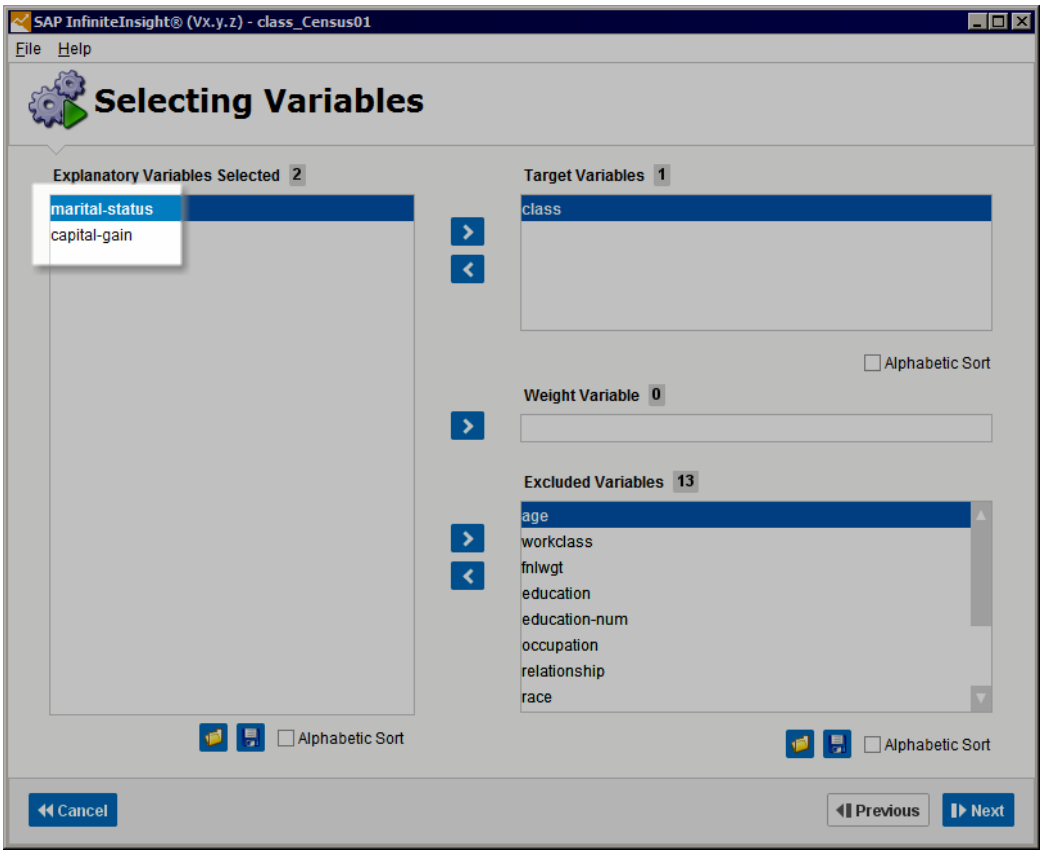

SAP InfiniteInsight® (Vx.y.z) - class_Census01
File Help
Selecting Contributory Variables
Targets: class
Smart Selection
View Reports
Variable | Maximum Contribution | KI | KR | KI + ... | Selec... | r | Priority | Last Iteration
marital-status 25.61% 0.532 0.989 1.521 6
capital-gain 17.71% 0.18 0.988 1.167 6
occupation 13.42% 0.457 0.994 1.451 6
education-num 10.88% 0.436 0.985 1.421 6
age 7.29% 0.411 0.997 1.408 5
capital-loss 6.29% 0.068 0.999 1.067 3
hours-per-we... 5.79% 0.345 0.991 1.336 4
education 5.24% 0.436 0.985 1.421 2
relationship 3.01% 0.553 0.988 1.541 1
sex 2.12% 0.23 0.991 1.22 1
workclass 1.1% 0.176 0.99 1.166 1
native-country 0.73% 0.048 0.995 1.043 1
race 0.63% 0.069 0.994 1.063 1
fnlwgt 0.2% 0.038 0.995 1.033 1
Number of Selected Variables: 2
Cancel
Previous
Next

6 Click *Next*. A message "*This will reset the current. Do you really want to do this?*" appears.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

7 Click *Yes* to move to the screen *Selecting Variables*.



8 Resume the model configuration from the step Selecting Variables.

Generating the Source Code of a Model

The feature used to generate the source code of a model is InfiniteInsight® Scorer. For more information on the InfiniteInsight® Scorer feature read the Integration Guide for KMX Generated Codes.

Caution

A specific license is needed to use this feature.The code file generated by SAP InfiniteInsight® contains all information necessary to the model such as variable encodings, missing value replacement, compressions, model parameters.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Generate the Code Corresponding to the Model

1. In the list *Target to be used*, choose the target of model.
2. Use the list *Information to be generated* to select the type of results:

Selected Option	Results of the Generated Model
Score/Estimates	score value (classification) or estimates (regression)
Probability	score value and probability value, except for HTML and all SQL codes for which only the probability value is provided.
Bar	score value and error bar value, except for HTML and all SQL codes for which only the error bar value is provided.

Caution

Both options Probability and Bar are only available for InfiniteInsight® Modeler - Regression/Classification models with nominal targets.

Note

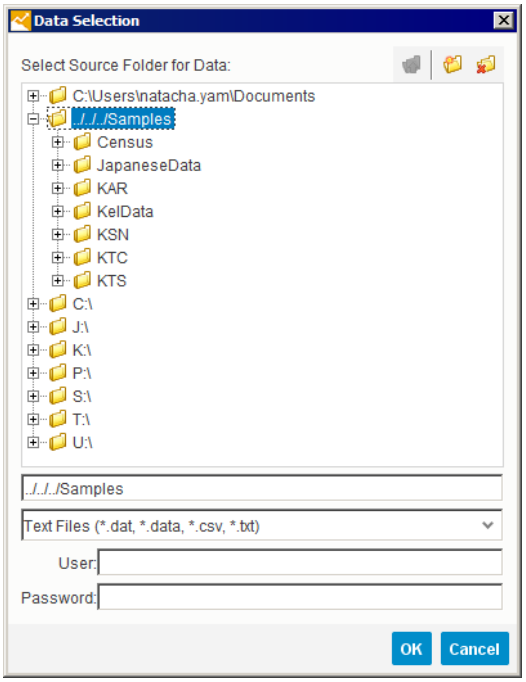
In the case of a continuous variable, the generated code (SQL for example) always includes a number of categories that is higher than in the user defined structure or as given by the parameter band count if no user has structure has been set. Indeed, the encoding of variables adds target curve points to increase the accuracy of coding according to the training data set. These points split some existing categories and thus increase the number of categories in the generated code.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

3 In the section *Code Settings*, select the code type to be generated (List of Generated Codes (on page 188)).

4 Click the *Browse* button associated with the *Folder* field and select a folder to save the generated file.

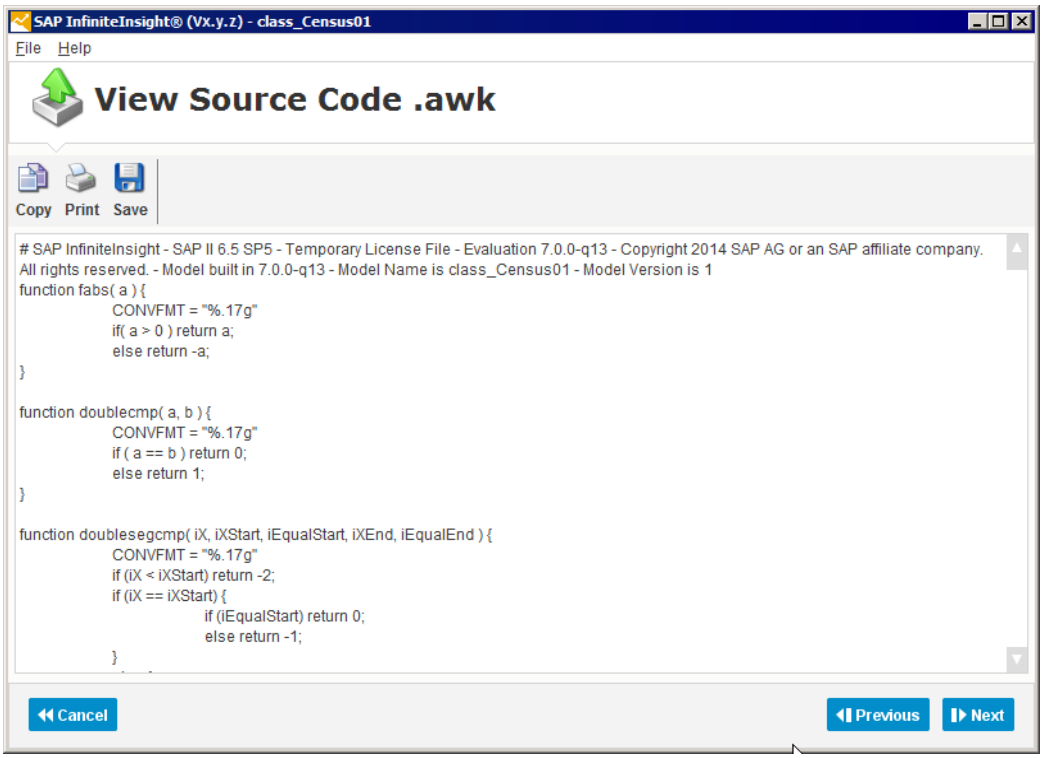
5 In the field *Generated File*, enter the name of the exported file. If you want to replace an existing file, use the *Browse* button to select it.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

6 If you have selected the option *View Generated Code*, it is displayed at the end of the generation process.

7 Click the *Generate* button.
The figure below shows the beginning of a sample C source code of an InfiniteInsight model.


SAP InfiniteInsight® (Vx.y.z) - class_Census01
File Help
View Source Code .awk
Copy
Print
Save
# SAP InfiniteInsight - SAP II 6.5 SP5 - Temporary License File - Evaluation 7.0.0-q13 - Copyright 2014 SAP AG or an SAP affiliate company. All rights reserved. - Model built in 7.0.0-q13 - Model Name is class_Census01 - Model Version is 1
function fabs(a) {
CONVFMT = "%.17g"
if(a > 0) return a;
else return -a;
}
function doublecmp(a, b) {
CONVFMT = "%.17g"
if (a == b) return 0;
else return 1;
}
function doublesegcmp(iX, iXStart, iEqualStart, iXEnd, iEqualEnd) {
CONVFMT = "%.17g"
if (iX < iXStart) return -2;
if (iX == iXStart) {
if (iEqualStart) return 0;
else return -1;
}
Cancel
Previous
Next

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

List of Generated Codes

The following table lists the available codes with their particularities:

Generated Code	Comment
C	see C Code Generator documentation
JAVA	needs the KxJRT.jar package to run
PMML3.2	
AWK	
CPP	
SAS	
SQLServer	wraps variable names with [ ]
SQLServerUDF	
HANA	InfiniteInsight® Scorer manages HANA column and row storage.
ORACLE	
OracleUDF	
SQLDB2	
DB2UDF	
DB2V9	
SQLTeradata	Teradata databases
TERAUDF	
MYSQL	
MYSQLUDF	
SybaseIQ	
SybaseIQUDF	
SQLNetezza	
SQLVertica	
PostgreSQL	
Geenplum	
Hive	
HTML (Javascript)	contains a form to fill which reproduces the InfiniteInsight® model
ScoreCard	only available for InfiniteInsight® Modeler - Regression/Classification models

Note

When generating SQL and SAS codes, you will be asked to provide the names of the key column and of the data set used.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Caution

Only `SQLServer` key code handles trimmed data during its execution. For other codes, if data are not trimmed it may generate some differences.

Advanced Settings

UNICODE Mode

The option *Activate UNICODE Mode* allows you to generate the code selected in Unicode so that it supports non-latin languages such as Japanese, Russian, and so on.

Note

This option is particularly useful for SQL codes.

SQL/UDF Options

- The option *Do not generate code for non-contributive variables* allows you to exclude from the code all variables with a contribution of 0 since they do not influence the result. In some cases, this can significantly reduce the size of the generated code.
- You can either *Use the default separator* ("GO") or *Use a custom separator*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

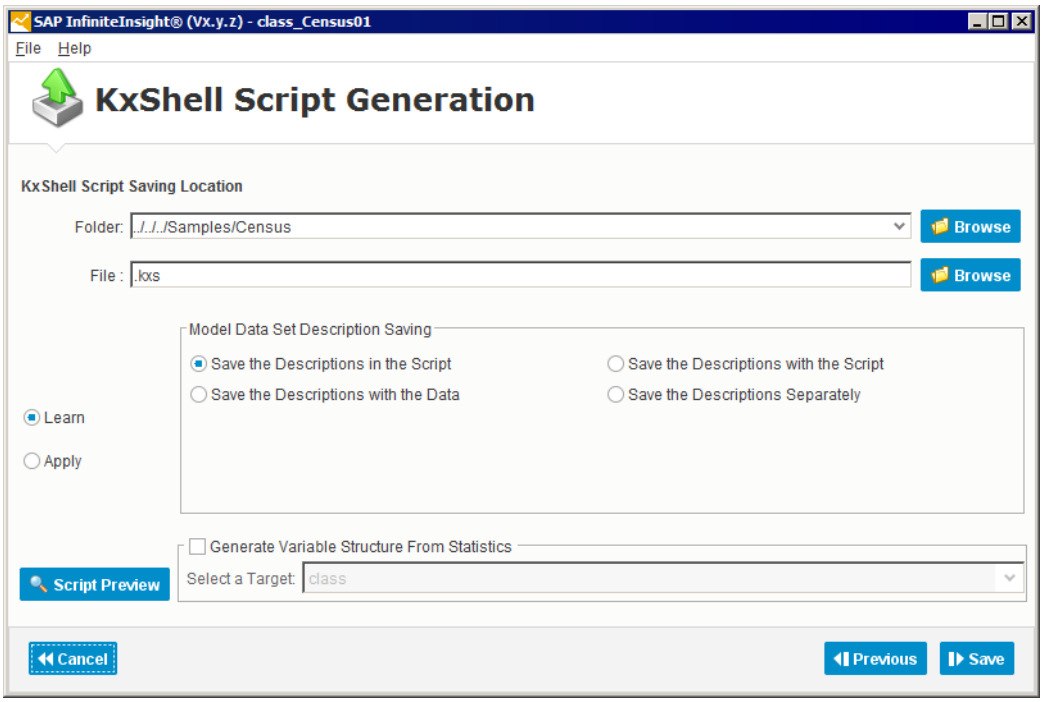
Exporting the Model as a KxShell Script

The KxShell script export allows you to generate a KxShell script reproducing the current model. This script can be used to run models in batches.

One easy way to get special settings in exported KxShell scripts is to first do the corresponding operation in the graphical user interface. For example, if you run an auto-selection of variables before exporting the shell script, then the exported script will include the code needed to do the auto-reduction.

☑ To Save the KxShell Script

1 In the section *Save/Export* of the menu *Using the Model*, select the option *Export KxShell Script*. The panel *KxShell Script Generation* is displayed.



2 Use the *Browse* button located to the right of the *Folder* field to select where the script will be saved.

3 In the field *KxShell Script*, enter the name of the file in which the script will be saved.

4 In the frame *Model Data Set Description Saving*, select where you want to save the data description. The four available options are:

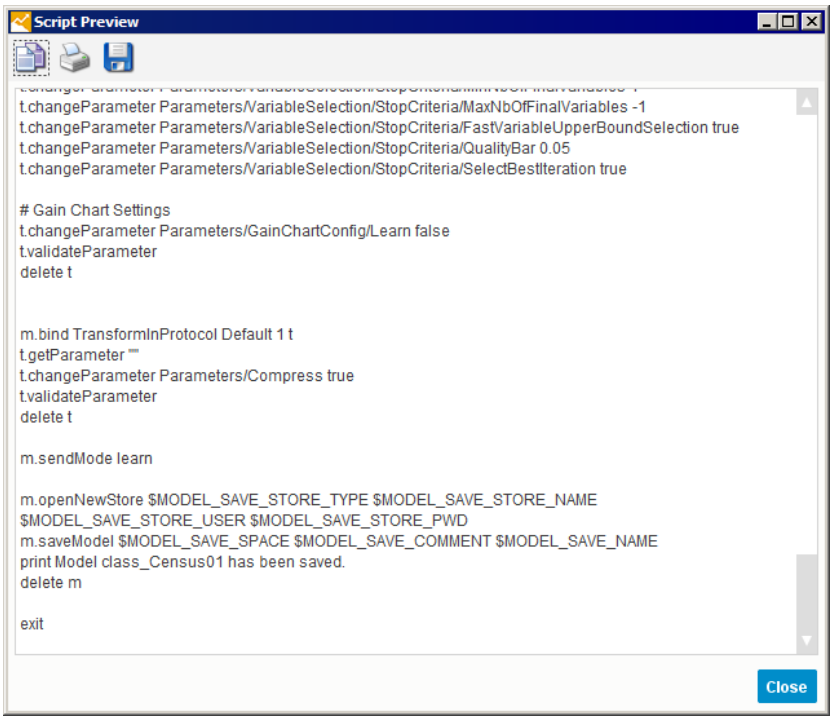
- *Save the Description in the Script*
  the data description is added in the KxShell script. Only one file is generated.
- *Save the Description with the Script*
  the data description is saved in an additional file in the same folder as the KxShell script
- *Save the Description with the Data*
  the data description is saved in an additional file in the same folder as data used for the model
- *Save the Description Separately*
  the data description is saved in an additional file. The user indicates the type of the description (text file, data base, flat memory, ...) and the location where the data description should be saved

> **Note**
>
> When saving the description in an additional file, the file is named following this syntax: KxDesc_<Dataset Role>_<Dataset Name>. For example, for a training data set named Census.csv, the description file name will be KxDesc_Training_Census.csv.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

5 Additionally you can export the variable structure with relation to a target variable by checking the option *Export Variable Structure in Script* and selecting the target variable in the list *Select a Target*. This option allows you to force the grouping of categories when training the model on new data sets.

6 Before exporting the script you can view the script by clicking the button *Script Preview*.



7 Click the *Next* button to start the generation process. Once the script has been generated, the menu *Using the Model* is displayed.

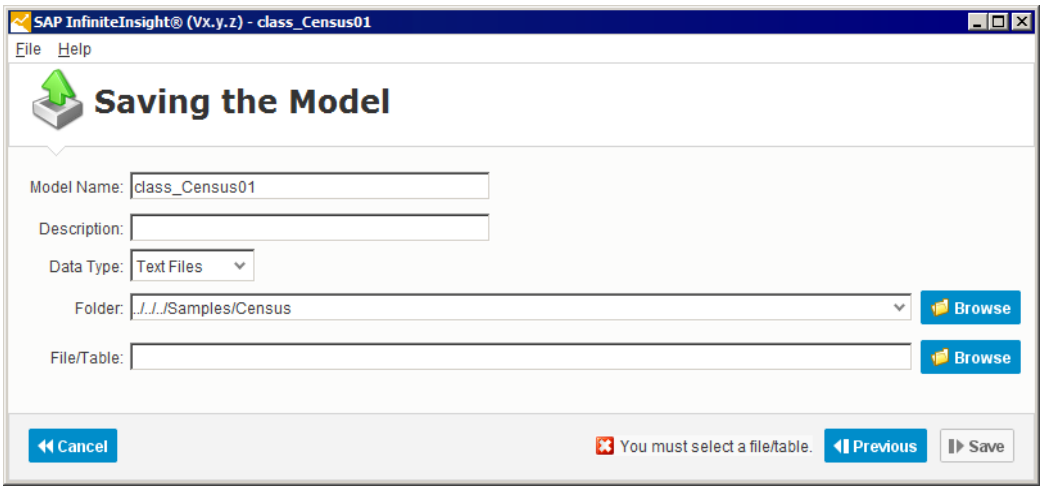
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Saving the Model

Once a model has been generated, you can save it. Saving it preserves all the information that pertains to that model, that is, the modeling parameters, its profit curves, and so on.

☑ To Save the Model

1 On the screen *Using the Model*, click the option *Save the Current Model*.
The screen *Saving the Model* appears.



2 Complete the following fields:

- *Model Name*: This field allows you to associate a name with the model. This name will then appear in the list of models to be offered when you open an existing model.
- *Description*: This field allows you to enter the information of your choosing, such as the name of the training data set used, the polynomial degree or the predictive power and prediction confidence obtained for the model. This information could be useful to you later for identifying your model.
- *Data Type*: this list allows you to select the type of storage in which you want to save your model. The following options are available:

- *Text files*, to save the model in a text file.
- *Database*, to save the model in a database.
- *Flat Memory*, to save the model in the active memory.
- *SAS Files*, to save the model in a SAS compatible file for a specified version of SAS and a specified platform (SAS v6 or 7/8 for Windows or UNIX).
- *SAS Transport*, to save the model in a generic SAS compatible file.
  - *Folder*: Depending upon which option you selected, this field allows you to specify the ODBC source, the memory store or the folder in which you want to save the model.
  - *File/Table*: This field allows you to enter the name of the file or table that is to contain the model. The name of the file must contain one of the following format extensions *.txt* (text file in which the data is separated by tabs) or *.csv* (text file in which the data is separated by commas).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Files Created when Saving a Model

When you save a model, SAP InfiniteInsight® creates a series of files in the specified store. The following table lists the files or tables created when saving a model and in which case.

Filename	Description	Used by
`KxAdmin`	lists all the models contained in the folder/database with additional information (date, version, name of the model, comments)	all models created with InfiniteInsight
`<Model_name>`	file named after the model and containing all the model data, except graphs information. Graphs are stored in additional tables (see below)	all models created with InfiniteInsight
`KxInfos`	indicates which additional tables are needed by the model	all models created with InfiniteInsight
`KxOlapCube`	stores the OLAP Cube used by the decision tree when the option InfiniteInsight® Modeler - Regression/Classification as Decision is activated	InfiniteInsight® Modeler - Regression/Classification models with decision tree
`KxLinks`	contains the links from the graphs of the model	InfiniteInsight® Social models only
`KxNodes`	lists all the nodes from all the graphs and their attributes	InfiniteInsight® Social models only
`KxCommunities`	matches the nodes to their communities, if the community detection was enabled	InfiniteInsight® Social models only

Caution

When sharing or sending a model, *all these files must be joined to the model* or the recipient will not be able to open the model.

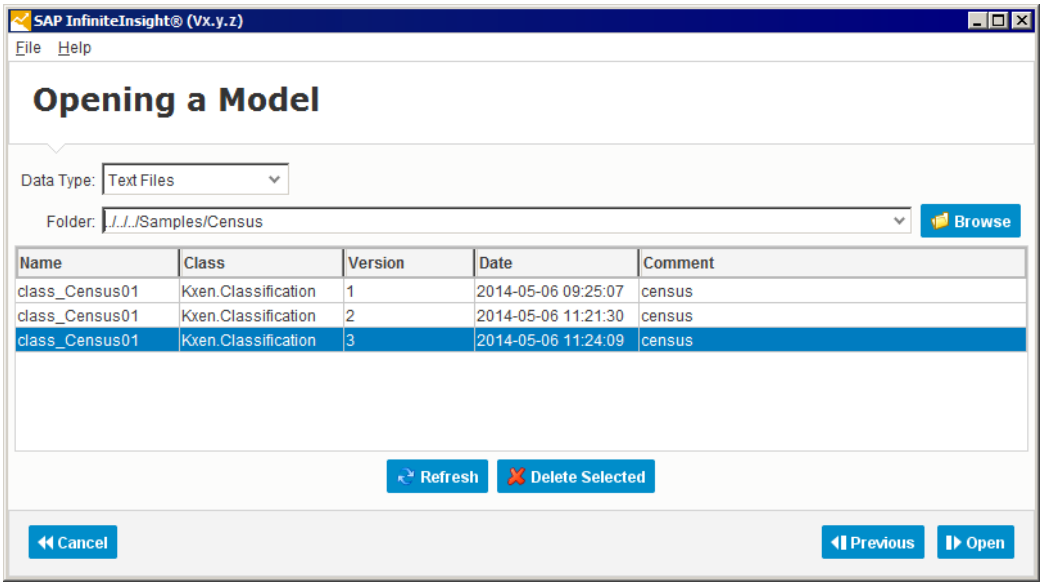
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Opening an Existing Model

Once saved, models may be opened and reused in SAP InfiniteInsight®.

☑ To Open a Model

1 On the main InfiniteInsight® modeling assistant screen, select *Load a Model*.
The screen *Opening a Model* appears.



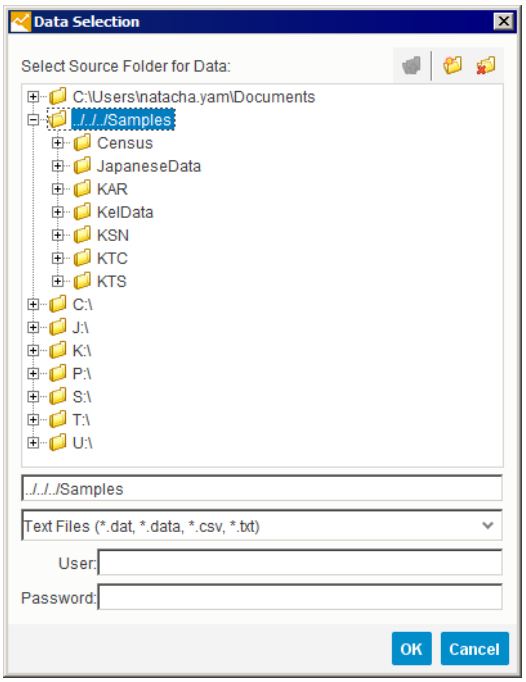
2 In the *Data Type* list, select one of the following options:

- *Text files*,
- *Database*,
- *SAS files*,
- *SAS Transport*,

depending upon the format of the model that you want to open.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

3 Click the *Browse* button.
A selection dialog box appears.



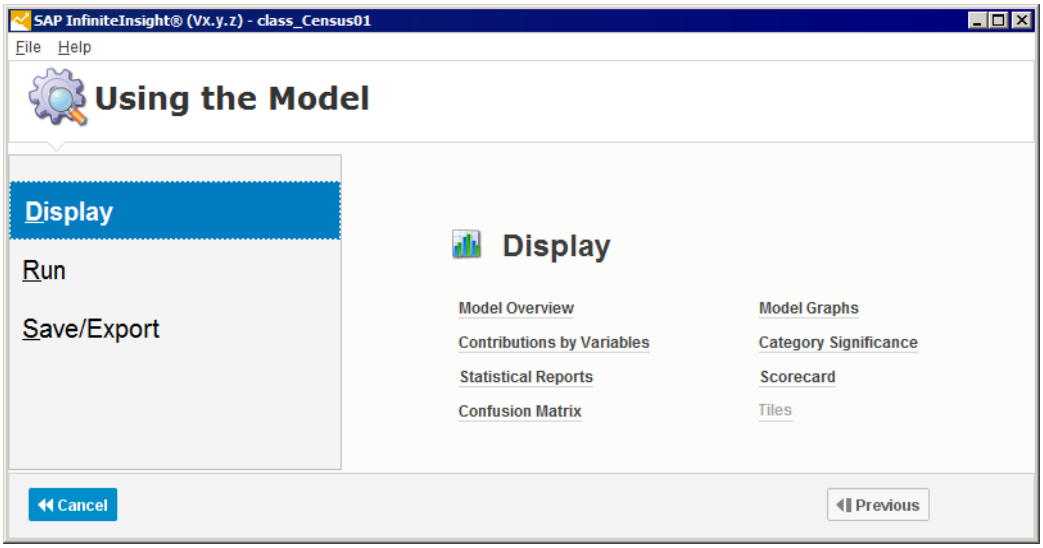
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

4 Select the folder that holds the model that you want to open.
The list of models contained in that folder appears providing the following information for each model.

Column	Description	Values
Name	Name under which the model has been saved	Character string
Class	Class of the model, that is the type of the model	▪ `Kxen.Classification`: Classification/Regression with nominal target ▪ `Kxen.Regression`: Classification/Regression with continuous target ▪ `Kxen.Segmentation`: Clustering with SQL Mode ▪ `Kxen.Clustering`: Clustering without SQL Mode ▪ `Kxen.TimeSeries`: Time Series ▪ `Kxen.AssociationRules`: Association Rules ▪ `Kxen.SimpleModel`: Classification/Regression and Clustering multi-target models, any other model
Version	Number of the model version when the model has been saved several times	Integer starting at 1
Date	Date when the model has been saved	Date and time in the format `yyyy-mm-dd hh:mm:ss`
Comment	Optional user defined comment that can be used to identify the model	Character string

5 Select a model from the list.

6 Click the *Open* button.
The screen **Using the Model** appears.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

6 InfiniteInsight® Modeler - Segmentation/Clustering

IN THIS CHAPTER

Application Scenario: Customize your Communications using Data Modeling 197
Creating a Clustering Model Using InfiniteInsight® Modeler 207

6.1 Application Scenario: Customize your Communications using Data Modeling

In this scenario, you are the Marketing Director of a large retail bank. The bank wants to offer a new financial product to its customers. Your project consists of launching a direct marketing campaign aimed at promoting this product.

In order to **customize the marketing messages from the bank and improve communication** with the various customers and prospects for this new product, the senior management of the bank asks you to build a segmentation model of the customers of this product.

Using **InfiniteInsight® Modeler - Segmentation/Clustering**, you can rapidly develop a descriptive model with the least possible cost. This model shows the **characteristic profiles** of the customers interested in your new product, and thus responds to your business issue and fulfills your objectives.

6.1.1 Presentation

This scenario develops logically from scenario 1.

In scenario 1, using the **InfiniteInsight® Modeler - Regression/Classification** (formerly known as K2R) feature, you managed to accomplish all the objectives of your first marketing campaign, meeting the deadlines and within the budget you were allowed.

In order to **customize the marketing messages from the bank and improve communication** with the various customers and prospects for this new product, the senior management of the bank now asks you to build a segmentation model of the customers of this product.

Using the **InfiniteInsight® Modeler - Segmentation/Clustering** (formerly known as K2S) feature, you can rapidly develop a descriptive model with the least possible cost. This model shows the **characteristic profiles** of the customers interested in your new product, and thus responds to your business issue and fulfills your objectives.

 © 2014 SAP AG or an SAP affiliate company. All rights reserved-

6.1.2 Your Objective

Consider the following case.

Using InfiniteInsight® Modeler - Regression/Classification, you have contacted the prospects most likely to be interested in your new financial product, and identified the ideal number of prospects to contact out of the entire database meeting the deadlines and within the budget you were allowed (see the <DocK2R>).

To **improve the rate of return of your campaign**, senior management asks you to:

- **Build a segmentation model of your customers**,
- **Analyze the characteristics of the identified clusters**,
- **Define customized communications for each cluster**.

The segmentation model in particular should allow you to distinguish customer clusters **by virtue of their propensity to purchase the new high-end savings product** proposed by your firm. You will optimize your understanding of your customers.

6.1.3 Your Approach

For organizational reasons, you want to define **five groups of customers**, or clusters, and describe the customer profiles for each of these groups.

To accomplish this project, you will use the sample of 50,000 people who responded to your first test, during the previous campaign.

This file corresponds to the sample file *Census01.csv*, provided with SAP InfiniteInsight® and described in the section **Introduction to Sample Files** (see page 60).

6.1.4 Your Business Issue

In your marketing database, you have:

- A list of 1,000,000 prospects,
- A list of 50,000 prospects (people selected during the test phase of your campaign), whose responses to the campaign are known. This sample thus constitutes a training data set. This sample, taken from the complete database, also exhibits some missing values.

Your business issue thus consists of:

- Rapidly building a segmentation model using the training data set (or sample). The clusters obtained will allow you to better understand the profiles of the individuals in your database as a function of their propensity to purchase.
- Then applying the segmentation model obtained from the training data to the entire list of prospects to determine which cluster each individual should belong.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

6.1.5 Your Solutions

To select the individuals to whom you should send a mailing, there are several solutions you can use:

- An **intuitive method**,
- A **classical statistical method** (K-means, both ascending and descending hierarchical segmentation models),
- The **InfiniteInsight method**.

Intuitive Method

This method consists of using your knowledge of the various profiles exhibited by your customers. Thanks to the domain-specific knowledge that you have of your customers, you determine the criteria of the segmentation model intuitively, and build the clusters yourself.

The main disadvantage of this method is that the number of information items available for each customer will invariably grow with time. The more data your database accumulates, the harder it is for you to manually create clusters that take all data into consideration and to develop a response to your business issue. Furthermore, as the increasing volume of information requires you to build segmentation models with increasing frequency, the time required to build these segmentation models becomes increasingly more significant.

Finally, management may want you to rationalize your methods, and to perform your segmentation using a method not based purely on your intuition. Defending your segmentation method based on intuition may be difficult.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Classical Statistical Method

On the basis of the information that you have, a Data Mining expert could build a segmentation model. In other words, you could ask a statistical expert to create a mathematical model that would allow you to build clusters based on the profiles of your customers.

To implement this method, the statistician must:

- Perform a detailed analysis of your database.
- Prepare your database down to the smallest detail, specifically, encoding the variables as a function of their type (nominal, ordinal or continuous) in preparation for segmentation. The encoding strategy used will determine the type of segmentation model obtained. At this step, the statistician may unconsciously bias the results.
- Test different types of algorithms (K-means, both ascending and descending hierarchical segmentation models) and select the one best suited to your business issue.
- Evaluate the relevance of the clusters obtained, in particular, the response to your domain-specific business issue.

After a few weeks, the statistical expert will be able to provide a certain number of clusters, or homogeneous groups, to which each of the individuals of your database are assigned.

This method presents significant constraints. You must:

- Ensure that your statistical expert, who is usually from an external department, is available for the scheduled period,
- Ensure that the modeling costs will fit into your budget,
- Spend time explaining your domain-specific business issue to the statistician,
- Spend time understanding the results that are provided,
- Ask a programmer to write a program to determine the cluster associated with any new individual added to your database.

In addition, this method is not systematic. Two statisticians performing this segmentation on the same data set could obtain different results.

InfiniteInsight® Method

InfiniteInsight® Modeler - Segmentation/Clustering allows you to build a segmentation model of your customers in a few minutes, taking into consideration the interest expressed by your customers in your new product.

InfiniteInsight® Modeler - Segmentation/Clustering automatically detects interactions between the variables to build homogeneous sub-sets, or clusters. Each cluster is homogeneous with respect to the entire set of variables, and in particular with respect to the target variable, that is, for example, "responded positively to my test".

You will discover the characteristics of different clusters, such as those clusters with an excellent response rate and those with a poor response rate. In addition, if your customer database contains customer expenditures on your other products, you will also obtain information on product sale synergies, by cluster.

Using InfiniteInsight® Modeler - Segmentation/Clustering, you have access to all the analytical features needed to define the type of message to be sent to the cluster for each customer. You have homogeneous clusters that will allow you to respond to your business issue. Of particular importance, this segmentation is systematic: the results obtained do not represent a particular point of view of your data, and is robust or consistent. Two people performing this segmentation using the InfiniteInsight method would obtain the same results.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

6.1.6 Introduction to Sample Files

This guide is accompanied by the following sample data files:

- A data file *Census01.csv*
- The corresponding description file *desc_census.csv*.

These files allow you to evaluate SAP InfiniteInsight® features and take your first steps in using it.

Census01.csv is the sample data file that you will use to follow the scenarios of **InfiniteInsight® Modeler - Regression/Classification** and **InfiniteInsight® Modeler - Segmentation/Clustering**. This file is an excerpt from the American Census Bureau database, completed in 1994 by Barry Becker.

Note

For more information about the **American Census Bureau**, see "http://www.census.gov".

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

This file presents the data on 48,842 individual Americans, of at least 17 years of age. Each individual is characterized by 15 data items. These data, or variables, are described in the following table.

Variable	Description	Example of Values
age	Age of individuals	Any numerical value greater than 17
workclass	Employer category of individuals	▪ *Private* ▪ *Self-employed-not-inc*
fnlwgt	Weight variable, allowing each individual to represent a certain percentage of the population	Any numerical value, such as "0", "2341" or 205019".
education	Level of study, represented by a schooling level, or by the title of the degree earned	▪ *11th* ▪ *Bachelors*
education-num	Number of years of study, represented by a numerical value	A numerical value between 1 and 16
marital-status	Marital status	▪ *Divorced* ▪ *Never-married*
occupation	Job classification	▪ *Sales* ▪ *Handlers-cleaners*
relationship	Position in family	▪ *Husband* ▪ *Wife*
race	Ethnicity	▪ *White* ▪ *Black*
sex	Gender	▪ *Male* ▪ *Female*
capital-gain	Annual capital gains	Any numerical value
capital-loss	Annual capital losses	Any numerical value
native country	Country of origin	▪ *United States* ▪ *France*
class	Variable indicating whether or not the salary of the individual is greater or less than $50,000	▪ "1" if the individual has a salary of greater than $50,000 ▪ "0" if the individual has a salary of less than $50,000

Note

In order to avoid complicating the InfiniteInsight® Modeler - Regression/Classification and InfiniteInsight® Modeler - Segmentation/Clustering application scenarios, the variable *fnlwgt* is used as a regular explanatory variable in these scenarios, and not as a weight variable.

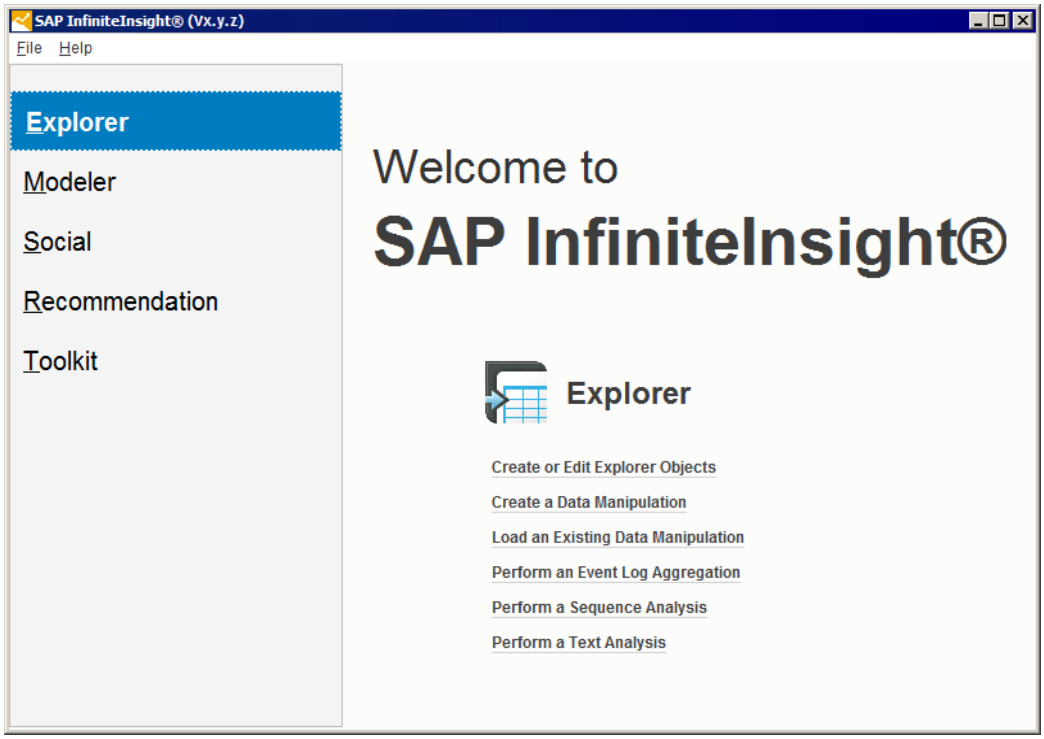
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

6.1.7 InfiniteInsight® modeling assistant

To accomplish the scenario, you will use the Java-based graphical interface of SAP InfiniteInsight®. This interface allows you to select the SAP InfiniteInsight® feature with which you will work, and help you at all stages of the modeling process.

☑ To Start InfiniteInsight® modeling assistant

1 Select *Start* > *Programs* > *SAP Business Intelligence* > *SAP SAP InfiniteInsight®* > *InfiniteInsight® modeling assistant.*
InfiniteInsight® modeling assistant screen appears.



2 Click the feature you want to use.

Editing the Options

☑ To Edit the Options of InfiniteInsight® modeling assistant

In the *File* menu, click *Preferences...*.
The window *Edit Options...* appears.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

The following options can be modified:

Category	Options
General	▪ Country ▪ Language ▪ Message Level ▪ Log Maximum Size ▪ Message Level for Strange Values ▪ Display the Parameter Tree ▪ Number of Store in the History ▪ Always Exit without Prompt ▪ Include Test in Default Cutting Strategy
Stores	▪ Default Store for Apply-in Data Set ▪ Default Store for Apply-out Data Set ▪ Default Store to Save Models
Metadata Repository	▪ Enable Single **Metadata Repository** ▪ Edit Variable Pool Content
Graphic	▪ Profit Curve Points ▪ Bar Count Displayed ▪ No InfiniteInsight® Lo*OK* and Feel ▪ Display 3D Chart ▪ Disable Double Buffering ▪ Optimize for Remote Display ▪ Remember Size and Position when Leaving
Report	▪ Number of Variables of Interest ▪ Active Style Sheet ▪ Customize Style Sheets

Customizing Style Sheets

SAP InfiniteInsight® offers the possibility to customize the generated reports. The default style sheet, called *SAP InfiniteInsight® Report Style Sheet (default)*, cannot be modified. You have to create your own style sheets to modify the settings.

Note

To create, load or save a style sheet, you have to indicate a data source in the panel *Edit Options...* before opening the window *SAP InfiniteInsight® Report Style Sheet Editor*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Create a New Style Sheet

1. In the field *Folder*, click the button (*Browse*).
2. Select a folder.
   This folder is your style sheets repository.
3. Click the button (*Add*).
   A new style sheet has been created.
4. Click the button .
   The panel *Report Style Sheet Editor* opens.
5. In the field *Style Sheet Name*, enter a name for the new style sheet.
   The extension `.krs` is automatically added.

Note

You can duplicate a style sheet by changing the name of your style sheet. The previous one is not deleted.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Delete a Style Sheet

1. Select one of the displayed style sheets.
2. Click the button (*Remove*). Note that the style sheet is not only deleted from the list but also from the data source.

☑ To Edit the General Settings

Settings...	Options...	Note...
Reports Background Color	▪ choose a color ▪ make transparent	Only the PDF and HTML formats can display a background color.
Edit Configuration	▪ font size ▪ font style ▪ font color ▪ text background color ▪ table configuration	Check the option *Dynamically render option changes* or click *Apply* when editing the settings, so that you can visualize the result.

The selected settings will be applied to both the wizard and the generated reports.

☑ To Edit the Charts Settings

Settings...	Options...	Note...
Chart Colors	▪ modify the charts colors	
Default Chart Bars Orientation	▪ horizontal ▪ vertical	It is possible to set another default orientation for specific report items.

☑ To Edit Report Items

1. Set the properties of your choice.
2. Click *Save* to validate.
   A window opens, indicating that your style sheet has been successfully saved.
3. Click *OK*.

Properties...	Functions...
Displayed as	name of the label
View Type	choose between Tabular, HTML and Graphical. The last one is only available if the report item can be displayed as a graph.
Chart Type	select one of the proposed chart types Note that this option is only available for report items of the view type *Graphical*.
Switch Bar Orientation	this option allows having another bar orientation as the default one for a specific report item
Sort by/Sort Order	you can select a column to sort by and choose between an ascending or a descending order
Visibility	you can hide columns of a report item or even menu items Note that at least one column of a report item must remain visible.

☑ To Apply the New Style Sheet to the Generated Reports

1. In the panel *Report*, select the new style sheet.
2. Click *OK*.
   A window opens, indicating that you have to restart the modeling assistant to take the edited options into account.
3. Click *OK*.
   When training a model, all the generated reports (the learn/excel/statistical reports) are now customized.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

6.2 Creating a Clustering Model Using InfiniteInsight® Modeler

Data modeling with InfiniteInsight® Modeler - Segmentation/Clustering is subdivided into four broadly defined stages:

1. Defining the Modeling Parameters
2. Generation and Validation of the Model
3. Analysis and Understanding of the Analytical Results
4. Using a Generated Model

6.2.1 Step 1 - Defining the Modeling Parameters

To respond to your business issue, you want to:

- **Break down** the sample of 50,000 prospects who responded to the test phase of your marketing campaign into homogeneous groups (see Summary of the InfiniteInsight® Modeler - Regression/Classification Application Scenario).
- **Describe** each of these groups and provide customized communication for each of these different groups.

The InfiniteInsight® Modeler - Segmentation/Clustering feature allows you to create descriptive models.

The first step in the modeling process consists of defining the modeling parameters:

1. Select a **data source** to be used as a training data set.
2. **Describe the selected data set** (on page 210).
3. Select the **variables**.
4. Select the **explanatory variables**.
5. **Checking the Modeling Parameters**.
6. Define the **number of clusters**. This step is optional.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Selecting a Data Source

For this Tutorial

Use the file *Census01.csv* as a training data set.

This file represents the sample that you had extracted from your database and used for the test phase of your direct marketing campaign. As specified in your test plan, this file contains 50,000 prospects, of which you know the behavior with respect to the new financial product:

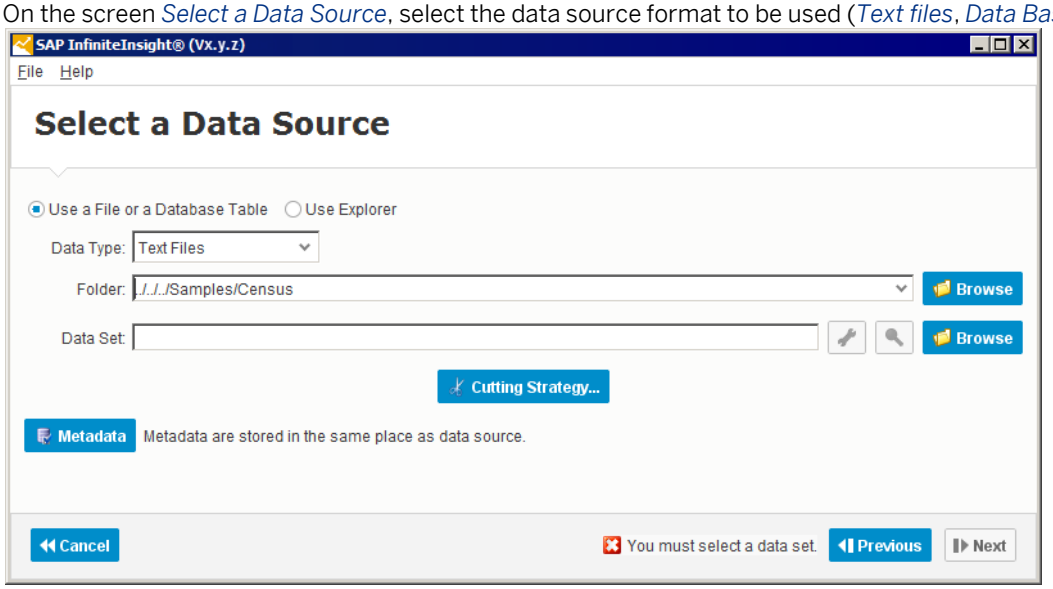
- 25% of the prospects showed themselves to be clearly interested. They chose to accept a meeting with one of your sales agents,
- 75% of the prospects declined your invitation.

In this file, you created a new variable Class, which corresponds to the reaction of prospects contacted during the test. You assigned:

- The value "1" for those prospects who responded positively to your invitation,
- The value "0" for those prospects who responded negatively to your invitation.

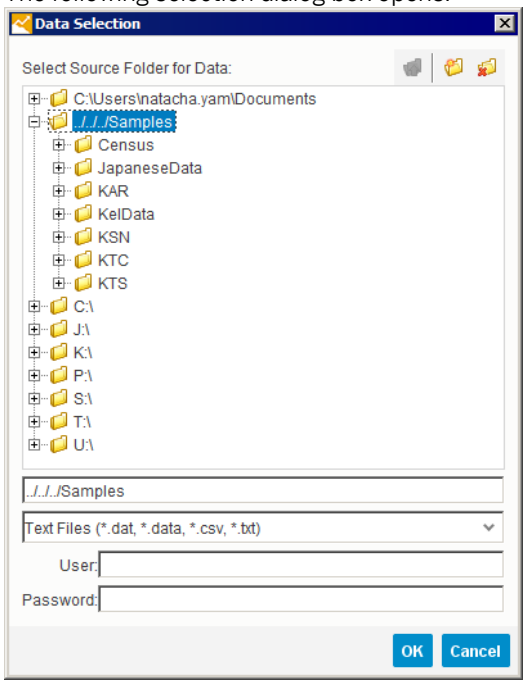
To Select a Data Source

1 On the screen *Select a Data Source*, select the data source format to be used (*Text files*, *Data Base*, ...).



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2 Click the *Browse* button.
The following selection dialog box opens.

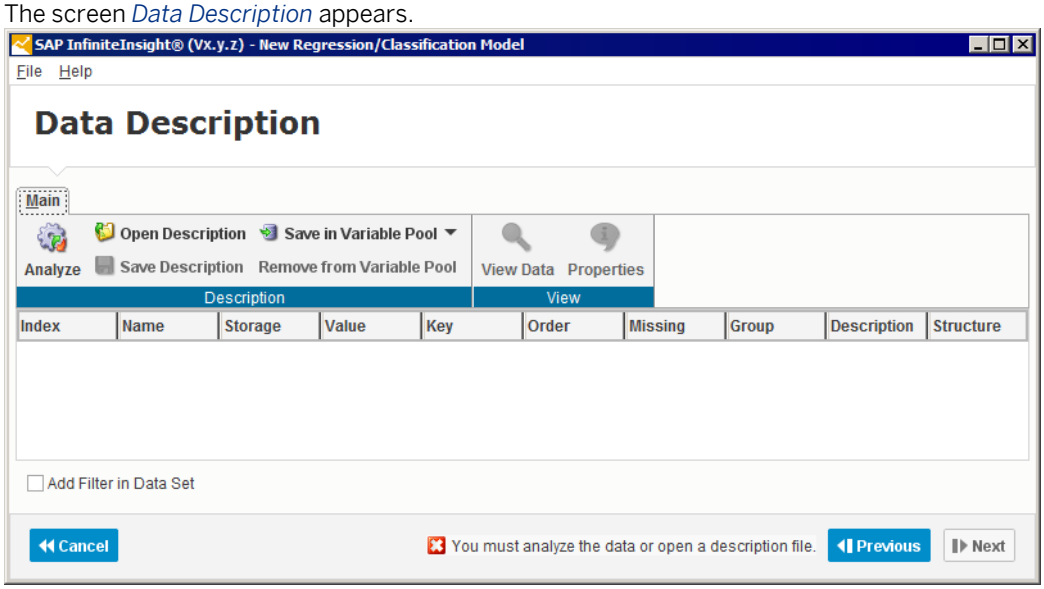


3 Double-click the *Samples* folder, then the *Census* folder.

> **Note**
>
> Depending on your environment, the *Samples* folder may or may not appear directly at the root of the list of folders. If you selected the default settings during the installation process, you will find the *Samples* folder located in *C:\Program Files\SAP InfiniteInsight\InfiniteInsightVx.y.z*

4 Select the file *Census01.csv*, then click *OK*.
The name of the file appears in the *Estimation* field.

5 Click *Next.*
The screen *Data Description* appears.



6 Go to section **Describing the Data Selected**.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

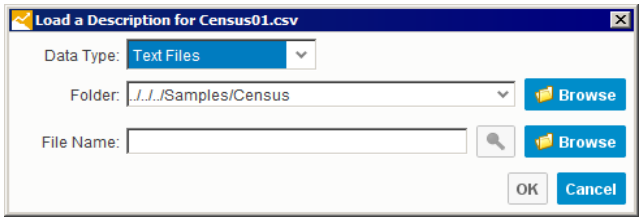
Describing the Data Selected

For this scenario

- Select *Text Files* as the file type.
- Use the file *Desc_Census01.csv* as the description file for the *Census01.cvs* data file.

To Select a Description File

1 On the screen *Data Description*, click the button *Open Description*. The following window opens:



2 In the window *Load a Description*, select the type of your description file.

3 In the *Folder* field, select the folder where the description file is located with the *Browse* button.

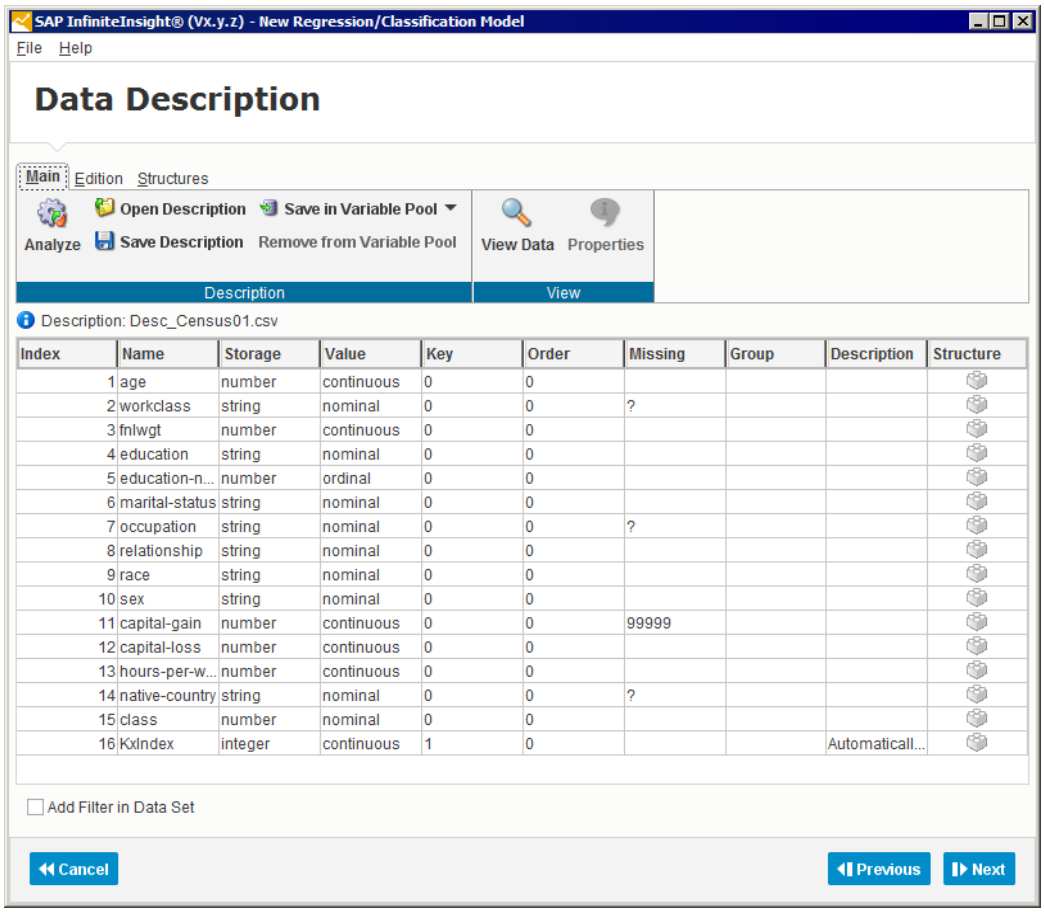
> **Note**
>
> The folder selected by default is the same as the one you selected on the screen *Data to be Modeled*.

4 In the *Description* field, select the file containing the data set description with the *Browse* button.

> **Caution**
>
> When the space used for model training **contains** a physical variable named `KxIndex`, it is not possible to use a description file without any key for the described space.
> When the space used for model training **does not contain** a physical variable named `KxIndex`, it is not possible to use a description file including a description about a KxIndex variable since it does not exist in current space.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

5 Click *OK*. The window *Load a Description* closes and the description is displayed on the screen *Data Description*.



Index	Name	Storage	Value	Key	Order	Missing	Group	Description	Structure
1	age	number	continuous	0	0				
2	workclass	string	nominal	0	0	?			
3	fnlwgt	number	continuous	0	0				
4	education	string	nominal	0	0				
5	education-n...	number	ordinal	0	0				
6	marital-status	string	nominal	0	0				
7	occupation	string	nominal	0	0	?			
8	relationship	string	nominal	0	0				
9	race	string	nominal	0	0				
10	sex	string	nominal	0	0				
11	capital-gain	number	continuous	0	0	99999			
12	capital-loss	number	continuous	0	0				
13	hours-per-w...	number	continuous	0	0				
14	native-country	string	nominal	0	0	?			
15	class	number	nominal	0	0				
16	KxIndex	integer	continuous	1	0			Automaticall...	

6 Click *Next*.

Why Describe the Data Selected?

In order for InfiniteInsight® features to interpret and analyze your data, they must be described. To put it another way, the description file must specify the nature of each variable, determining their:

- **Storage format**: number (*number*), character string (*string*), date and time (*datetime*) or date (*date*).
- **Type**: *continuous*, *nominal, ordinal* or *textual*.

For more information on data description, go to **Types of Variables** (on page 26) and **Storage Formats** (on page 29).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

How to Describe Selected Data

To describe your data, you can:

- Either **use an existing description file**, that is, taken from your information system or a previously created description file from InfiniteInsight® features,
- Or **create a description file using the Analyze option** from InfiniteInsight® modeling assistant. In this case, you must validate the description file obtained. You can save this file for later use.

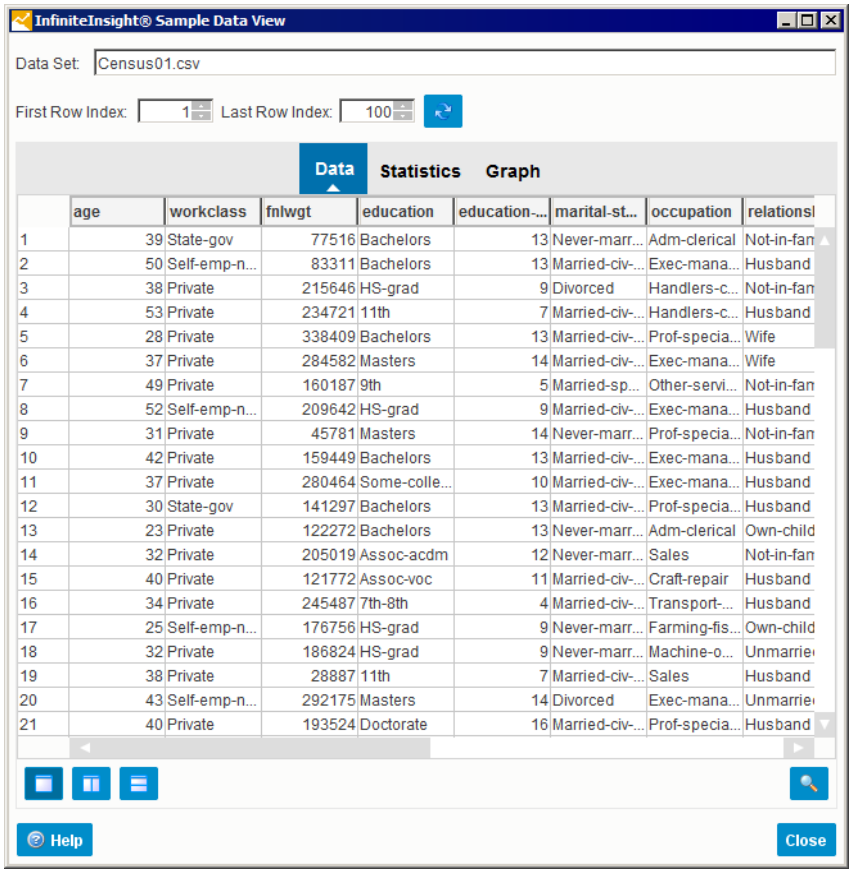
> **Caution**
>
> The description file obtained using the *Analyze* option results from the analysis of the first 50 lines of the initial data file. In order to avoid all bias, we encourage you to randomly sort your data set outside SAP InfiniteInsight® before performing this analysis.

Viewing the Data

To help you validate the description when using the *Analyze* option, you can display the first hundred lines of your data set.

☑ To View the Data

1 Click the button *View Data*. A new window opens displaying the data set top lines:



	age	workclass	fnlwgt	education	education-...	marital-st...	occupation	relationsl
1	39	State-gov	77516	Bachelors	13	Never-marr...	Adm-clerical	Not-in-fan
2	50	Self-emp-n...	83311	Bachelors	13	Married-civ-...	Exec-mana...	Husband
3	38	Private	215646	HS-grad	9	Divorced	Handlers-c...	Not-in-fan
4	53	Private	234721	11th	7	Married-civ-...	Handlers-c...	Husband
5	28	Private	338409	Bachelors	13	Married-civ-...	Prof-specia...	Wife
6	37	Private	284582	Masters	14	Married-civ-...	Exec-mana...	Wife
7	49	Private	160187	9th	5	Married-sp...	Other-servi...	Not-in-fan
8	52	Self-emp-n...	209642	HS-grad	9	Married-civ-...	Exec-mana...	Husband
9	31	Private	45781	Masters	14	Never-marr...	Prof-specia...	Not-in-fan
10	42	Private	159449	Bachelors	13	Married-civ-...	Exec-mana...	Husband
11	37	Private	280464	Some-colle...	10	Married-civ-...	Exec-mana...	Husband
12	30	State-gov	141297	Bachelors	13	Married-civ-...	Prof-specia...	Husband
13	23	Private	122272	Bachelors	13	Never-marr...	Adm-clerical	Own-child
14	32	Private	205019	Assoc-acdm	12	Never-marr...	Sales	Not-in-fan
15	40	Private	121772	Assoc-voc	11	Married-civ-...	Craft-repair	Husband
16	34	Private	245487	7th-8th	4	Married-civ-...	Transport-...	Husband
17	25	Self-emp-n...	176756	HS-grad	9	Never-marr...	Farming-fis...	Own-child
18	32	Private	186824	HS-grad	9	Never-marr...	Machine-o...	Unmarrie
19	38	Private	28887	11th	7	Married-civ-...	Sales	Husband
20	43	Self-emp-n...	292175	Masters	14	Divorced	Exec-mana...	Unmarrie
21	40	Private	193524	Doctorate	16	Married-civ-...	Prof-specia...	Husband

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2 In the field *First Row Index*, enter the number of the first row you want to display.
3 In the field *Last Row Index*, enter the number of the last row you want to display.
4 Click the *Refresh* button to see the selected rows.

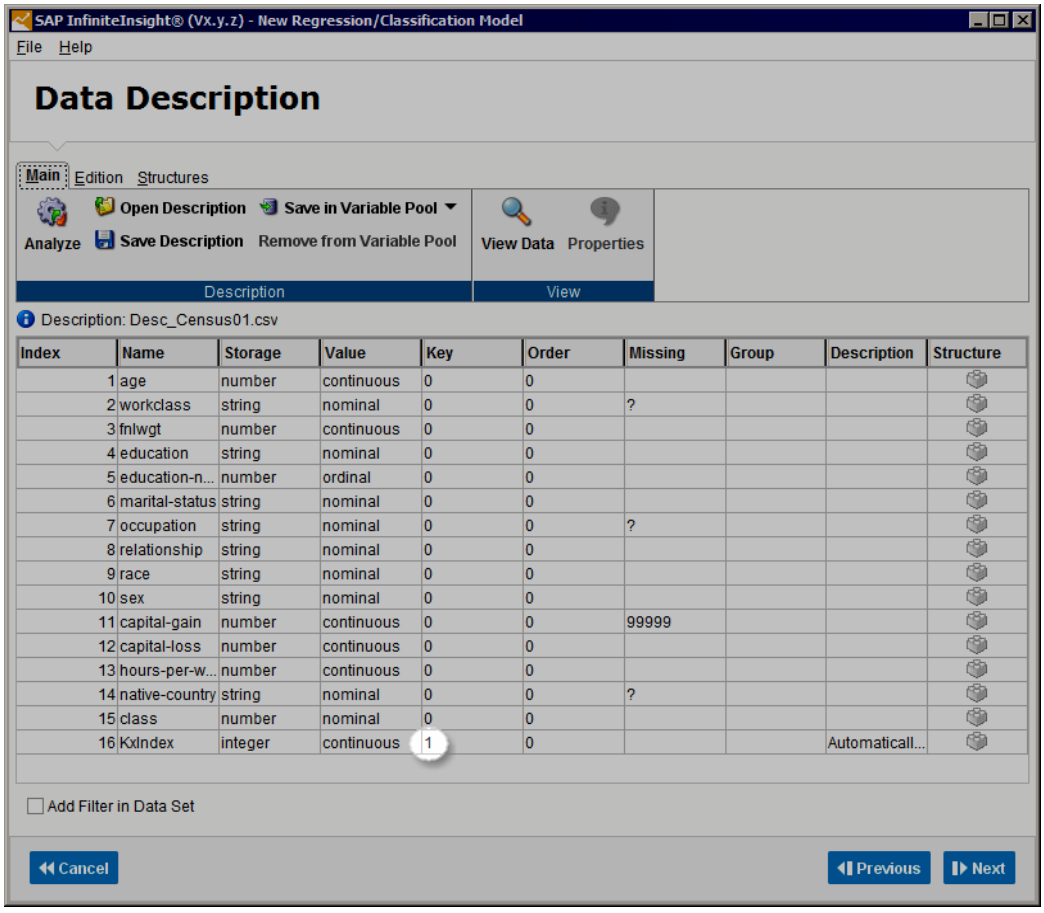
A Comment about Database Keys

For data and performance management purposes, the data set to be analyzed must contain a variable that serves as a key variable. Two cases should be considered:

- If the initial data set **does not contain a key variable**, a variable index *KxIndex* is automatically generated by SAP InfiniteInsight® features. This will correspond to the row number of the processed data.
- If the file **contains one or more key variables**, they are not recognized automatically. You must specify them manually in the data description. See the procedure To Specify that a Variable is a Key (see page 75).

To Specify that a Variable is a Key

1 In the *Key* column, click the box corresponding to the row of the key variable.
2 Type in the value "*1*" to define this as a key variable.



Index	Name	Storage	Value	Key	Order	Missing	Group	Description	Structure
1	age	number	continuous	0	0				
2	workclass	string	nominal	0	0	?			
3	fnlwgt	number	continuous	0	0				
4	education	string	nominal	0	0				
5	education-n...	number	ordinal	0	0				
6	marital-status	string	nominal	0	0				
7	occupation	string	nominal	0	0	?			
8	relationship	string	nominal	0	0				
9	race	string	nominal	0	0				
10	sex	string	nominal	0	0				
11	capital-gain	number	continuous	0	0	99999			
12	capital-loss	number	continuous	0	0				
13	hours-per-w...	number	continuous	0	0				
14	native-country	string	nominal	0	0	?			
15	class	number	nominal	0	0				
16	KxIndex	integer	continuous	1	0			Automaticall...	

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Defining a Variable Structure

There are three ways to define a variable structure:

- by first extracting the categories from the variable statistics, then editing or validating the suggested structure.
- by importing the structure from an existing model.
- by building a new structure from scratch.

The option *Optimal Grouping* allows you to let Data Encoding group together the categories groups defined in the variable structure if they bring the same information.

For more details on variable structure, see InfiniteInsight® Modeler - Regression/Classification > Defining a Variable Structure.

Filtering the Data Set

In order to accelerate the learn process and to optimize the resulting model, you can apply a filter to your data set.

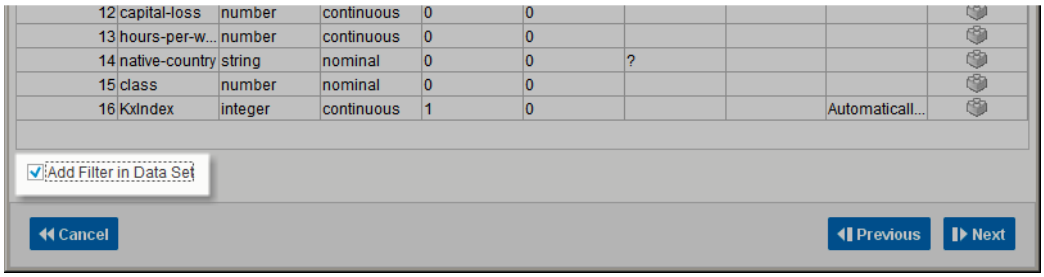
For this scenario

Do not use the filtering option.

To Filter a Data Set

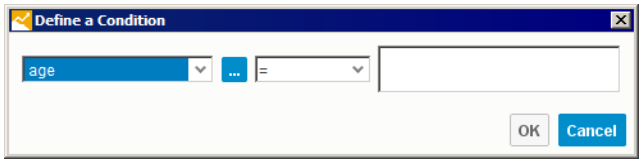
1 Check the option *Add a Filter in Data Set*.

2 Click *Next*.


12 capital-loss number continuous 0 0
13 hours-per-w... number continuous 0 0
14 native-country string nominal 0 0 ?
15 class number nominal 0 0
16 KxIndex integer continuous 1 0 Automaticall...
Add Filter in Data Set
Cancel
Previous
Next

To Add a Condition

1 Click the button *Add a Condition*.
The window *Define a Condition* opens.



2 Choose a variable in the first list.

3 Choose an operator in the second list.

4 Indicate a value in the third list:

- For a variable with number storage, type a value.
- For a variable with string storage, choose a variable in the list. If the list is empty click the button , to extract the variable categories.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

5 Click *OK*.

> **Note**
>
> You can edit a condition by double-clicking it.

☑ To Add a Logical Conjunction

Click the button *Add Logic 'And'* or the button *Add Logic 'Or'*.

> **Note**
>
> You can change a conjunction by double-clicking it.

☑ To Change the Order

You can change the order of the nodes to accelerate the filtering process by setting the conditions with the highest probability to be false at the top of the list.

1 Select the node you want to move up or down.

2 Use the buttons and to change its position in the filter.

☑ To Delete a Node

1 Select the node you want to delete.

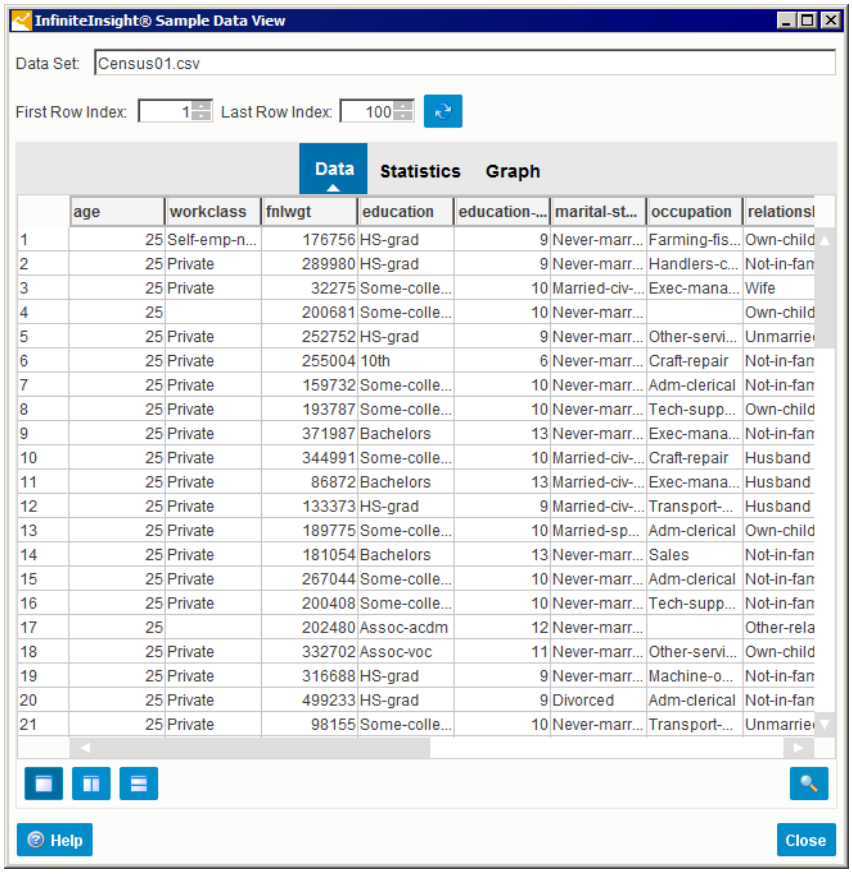
2 Click the button *Remove Selected Node*.

☑ To Display the Filtered Data Set

You can visualize the data set that you will obtain after the application of the filter.

Click the button *View Data*.
A pop-up window opens.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-



	age	workclass	fnlwgt	education	education-...	marital-st...	occupation	relationsl
1	25	Self-emp-n...	176756	HS-grad	9	Never-marr...	Farming-fis...	Own-child
2	25	Private	289980	HS-grad	9	Never-marr...	Handlers-c...	Not-in-fan
3	25	Private	32275	Some-colle...	10	Married-civ-...	Exec-mana...	Wife
4	25		200681	Some-colle...	10	Never-marr...		Own-child
5	25	Private	252752	HS-grad	9	Never-marr...	Other-servi...	Unmarrie
6	25	Private	255004	10th	6	Never-marr...	Craft-repair	Not-in-fan
7	25	Private	159732	Some-colle...	10	Never-marr...	Adm-clerical	Not-in-fan
8	25	Private	193787	Some-colle...	10	Never-marr...	Tech-supp...	Own-child
9	25	Private	371987	Bachelors	13	Never-marr...	Exec-mana...	Not-in-fan
10	25	Private	344991	Some-colle...	10	Married-civ-...	Craft-repair	Husband
11	25	Private	86872	Bachelors	13	Married-civ-...	Exec-mana...	Husband
12	25	Private	133373	HS-grad	9	Married-civ-...	Transport-...	Husband
13	25	Private	189775	Some-colle...	10	Married-sp...	Adm-clerical	Own-child
14	25	Private	181054	Bachelors	13	Never-marr...	Sales	Not-in-fan
15	25	Private	267044	Some-colle...	10	Never-marr...	Adm-clerical	Not-in-fan
16	25	Private	200408	Some-colle...	10	Never-marr...	Tech-supp...	Not-in-fan
17	25		202480	Assoc-acdm	12	Never-marr...		Other-rela
18	25	Private	332702	Assoc-voc	11	Never-marr...	Other-servi...	Own-child
19	25	Private	316688	HS-grad	9	Never-marr...	Machine-o...	Not-in-fan
20	25	Private	499233	HS-grad	9	Divorced	Adm-clerical	Not-in-fan
21	25	Private	98155	Some-colle...	10	Never-marr...	Transport-...	Unmarrie

☑ To Save a Filter

You can save the filter you created to be able to reuse it at a later moment without being obliged to recreate the same conditions.

1. Click the button *Save Filter*.
   A pop-up window is displayed.
2. In the list *Data Type*, select the format in which you want to save the filter.
3. Use the *Browse* button located on the right of the *Folder* field to select the folder or database where you want to save the filter.
4. In the *Description* field, enter the name of the file or table in which you want to save the filter.
5. Click *OK*.

☑ To Load an Existing Filter

To apply a filter to the data set, you can use a file created during a previous use of the data set in SAP InfiniteInsight®.

1. Click the button *Load Existing Filter*.
   A pop-up window is displayed.
2. Use the list *Data Type* to select the format of the filter.
3. Use the *Browse* button located on the right of the *Folder* field to select the folder or the database in which the filter is stored.
4. Use the *Browse* button located on the right of the *Description* field to select the file or the table containing the filter.
5. Click *OK*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Translating the Variable Categories

You can translate the categories of a nominal variable, save the translation or load an existing translation. This translation has no influence on the variable structure, which has to be set according to the original values of the variable.

> **Note**
>
> The variable "Target Key", that is used in the advanced settings for example, does not take into account the translation when displaying the possible values of this variable.

☑ To Translate the Variable Categories

1. Click a nominal variable to translate its categories.
2. Go to the *Edition* tab of the ribbon and click the option *Translate Categories*. A new window appears.
3. Choose into which languages you want to translate. By default, the language of the user interface is displayed as a column.
4. Click the button to extract the variable categories from the data set.
5. Translate the categories.

> **Note**
>
> You do not need to fill all fields.

1. Click *OK*.

☑ To Save the Categories Translation

1. Translate the variable categories as described above.
2. Click the *Save* button.
3. Choose a *Data Type*.
4. Select a *Folder*.
5. Enter a *Name* for the file or table.
6. Click *OK*.

☑ To Load an Existing Translation File

1. Click a nominal variable to translate its categories.
2. Go to the *Edition* tab of the ribbon and click the option *Translate Categories*. A new window appears.
3. Click the *Load* button.
4. Select the format of the translation in the list *Data Type*.
5. Use the *Browse* button located on the right of the *Folder* field to select the folder or the database in which the description is stored.
6. Use the *Browse* button located on the right of the field *Table or File* to select the file or the table containing the description.
7. Click *OK*.
8. Click the button *Update* to refresh the display of the categories.
9. If the list of columns is not named correctly, use the *Advanced Settings* (see next paragraph) to set a header line and update again.
10. Map the language names with those from the loaded translation, by clicking the categories and choosing the corresponding language in the contextual menu.
11. Click *OK*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Selecting Variables

Once the training data set and its description have been entered, you must select different variables :

- one or more targets variables,
  The InfiniteInsight® Modeler - Segmentation/Clustering feature is capable of segmenting a data set independently, that is, it does not require that a target variable be selected. However, even though this is not required, we strongly recommend selecting a target variable. For the process of segmenting, a data set gains maximum meaning only when it is accomplished in relation to a domain-specific business issue, expressed in the form of a target variable.
- possibly a weight variable,
- and the explanatory variables.

> **Note**
>
> For more information on variable roles, see section Role of Variables (see "Roles of Variables" on page 30).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

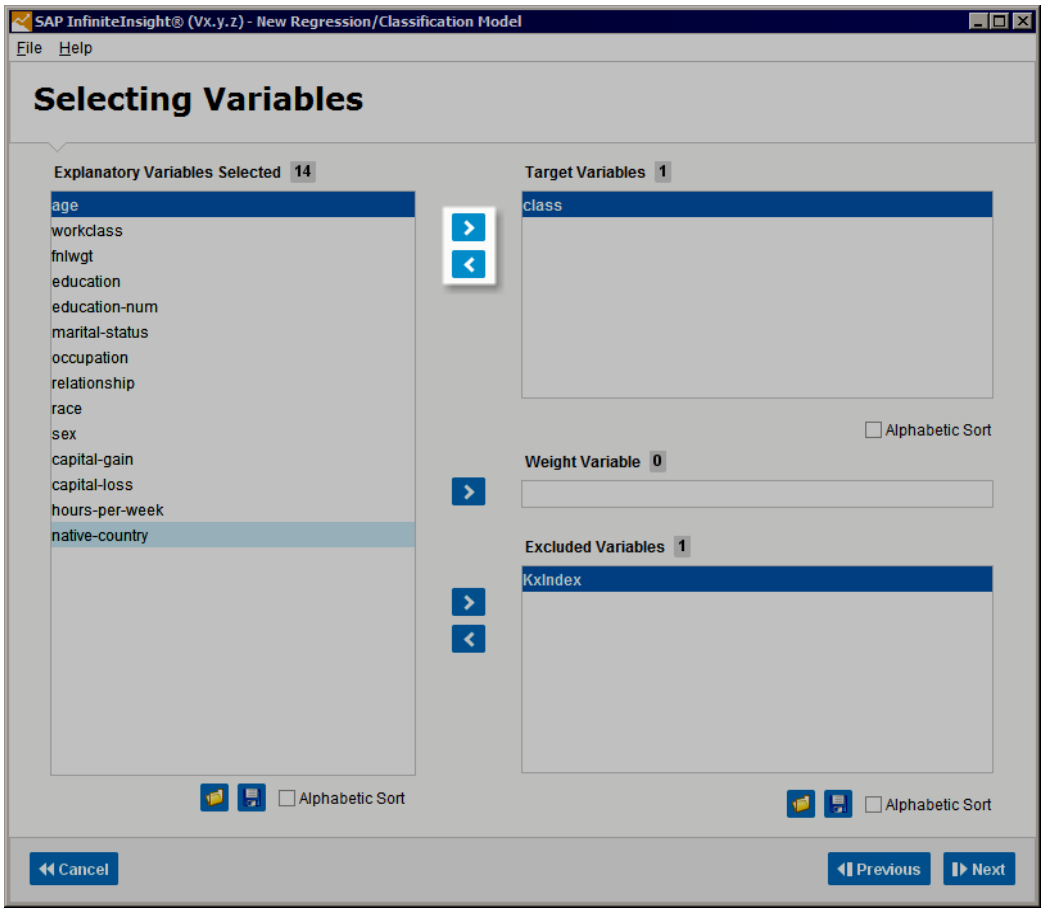
Target Variables

For this Scenario

Select the variable *Class* as your **target variable**, that is, the variable that indicates the probability of an individual responding in a positive or negative manner to your campaign.

To Select Targets Variables

1 On the screen *Selecting Variables*, in the section *Explanatory variables selected* (left hand side), select the variables you want to use as Target Variables.



> **Note**
>
> On the screen *Selecting Variables*, variables are presented in the same order as that in which they appear in the table of data. To sort them alphabetically, select the option *Alphabetic sort*, presented beneath each of the variables list.

2 Click the button > located on the left of the screen section *Target Variables* (upper right hand side).
The variable moves to the screen section *Target Variables*.
Also, select a variable in the screen section *Target Variables* and click the button < to move the variables back to the screen section *Explanatory variables selected*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

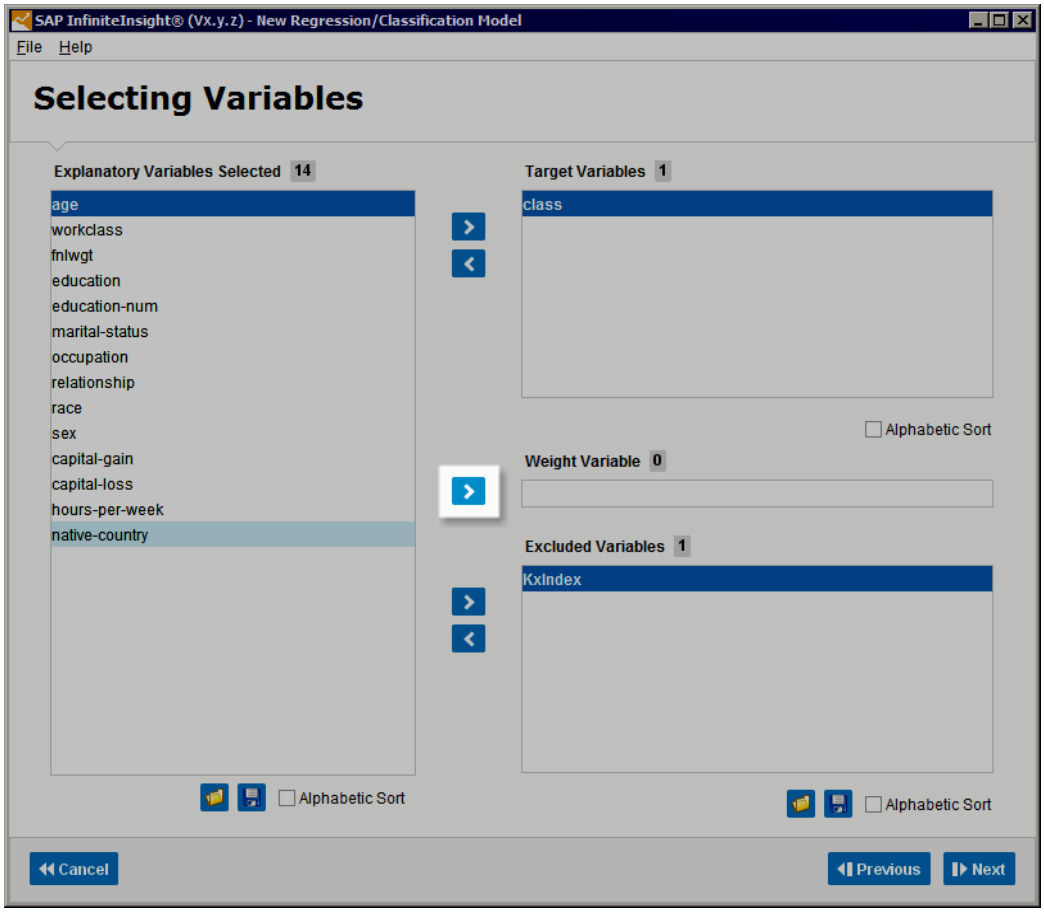
Weight Variable

For this Scenario

- Do **not** select a weight variable.

To Select a Weight Variable

1 On the screen *Selecting Variables*, in the section *Explanatory variables selected* (left hand side), select the variables you want to use as a Weight Variable.



> **Note**
>
> On the screen *Selecting Variables*, variables are presented in the same order as that in which they appear in the table of data. To sort them alphabetically, select the option *Alphabetic sort*, presented beneath each of the variables list.

2 Click the button > located on the left of the screen section *Weight Variable* (middle right hand side).
The variable moves to the screen section *Weight Variable*.
Also, select a variable in the screen section *Weight Variable* and click the button < to move the variables back to the screen section *Explanatory variables selected*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Explanatory Variables

By default, and with the exception of key variables, all variables contained in your data set are taken into consideration for generation of the model. You may exclude some of these variables.

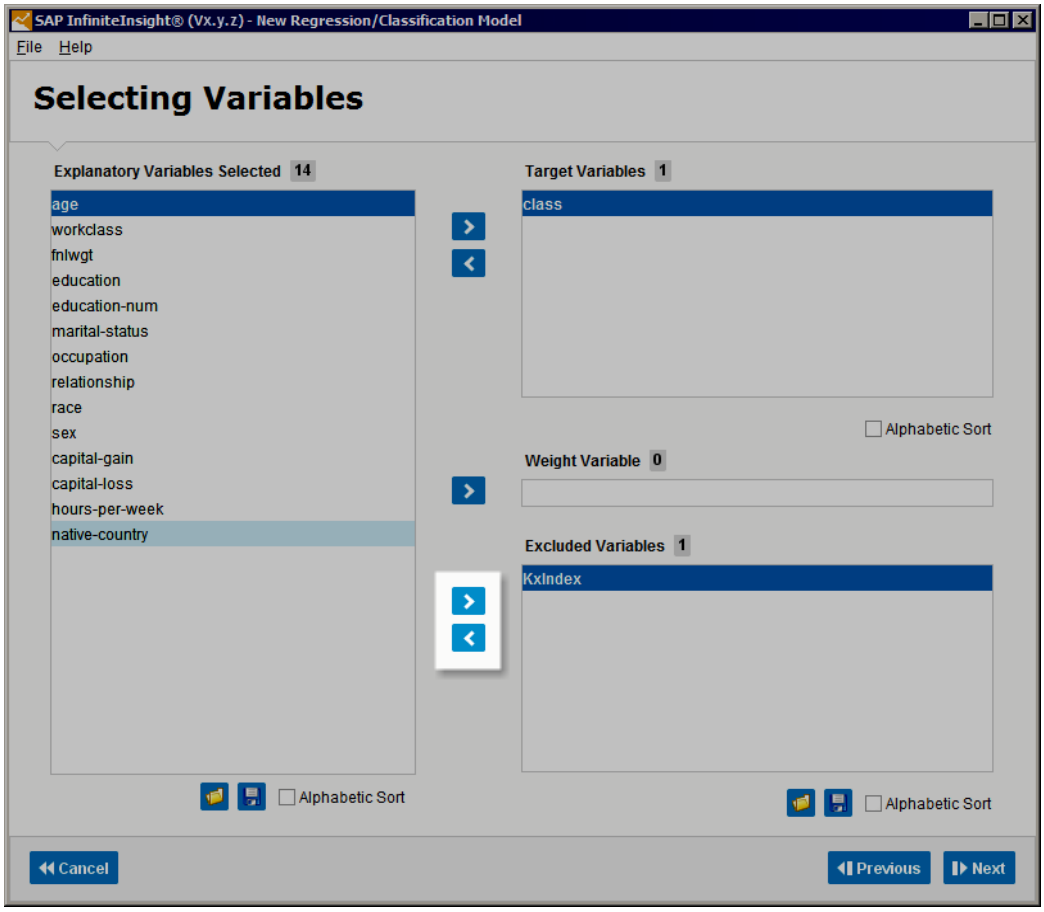
The decision whether to include or exclude a variable for generation of your segmentation model depends upon domain-specific considerations. Your domain-specific knowledge allows you to determine which variables are the most useful for description of the clusters or homogeneous groups. A regression model generated using InfiniteInsight® Modeler - Regression/Classification (*formerly known as K2R*) would also be used as a tool to determine the variables with the greatest explanatory power for a given phenomenon.

For this Scenario

- Exclude the variable *KxIndex*, as this is a key variable. Since the initial data set does not contain a key variable, SAP InfiniteInsight® feature generated KxIndex automatically.
- Retain all the other variables.

To Exclude some Variables from Data Analysis

1 On the screen *Selecting Variables*, in the section *Explanatory Variables Selected* (left hand side), select the variable to be excluded.



Note

On the screen *Selecting Variables*, variables are presented in the same order as that in which they appear in the table of data. To sort them alphabetically, select the option *Alphabetic sort*, presented beneath each of the variables list.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

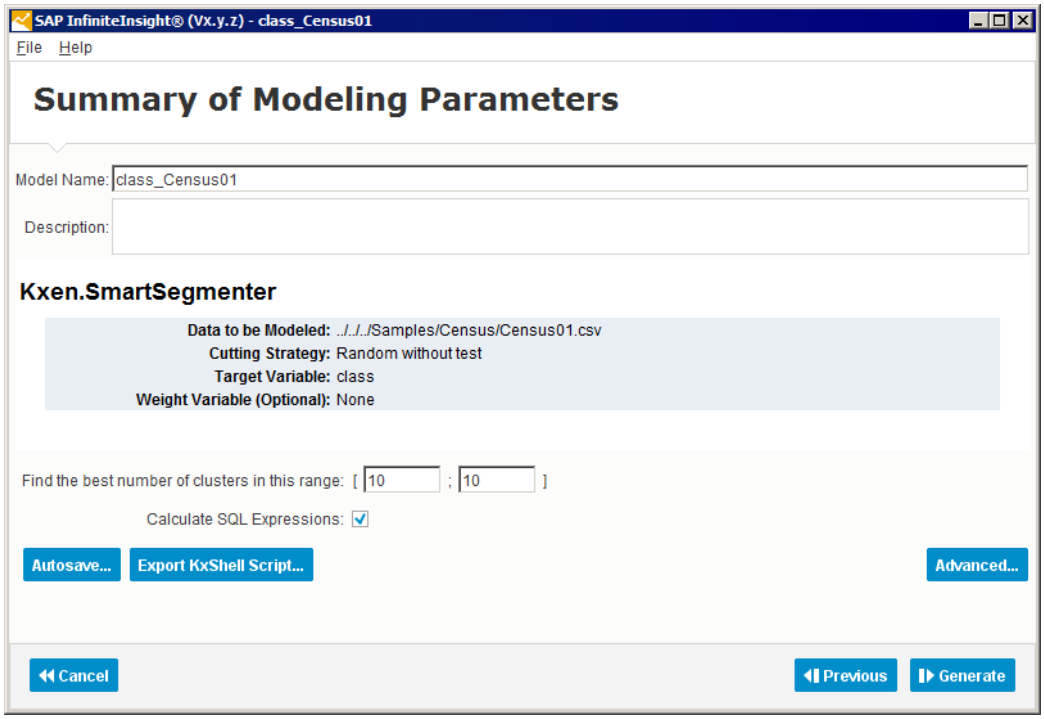
2 Click the button > located on the left of the screen section *Variables excluded* (lower right hand side).
The variable moves to the screen section *Variables excluded*.
Also, select a variable in the screen section *Variables excluded* and click the button < to move the variables back to the screen section *Explanatory variables selected*.

3 Click *Next*.
The screen *Summary of the Modeling Parameters* appears.

4 Go to the section **Checking Modeling Parameters**.

Checking Modeling Parameters

The screen *Summary of Modeling Parameters* allows you to check the modeling parameters just before generating the model.



Note

The screen *Summary of Modeling Parameters* contains an *Advanced* button. By clicking this button, you access the screen *Specific Parameters of the Model*. For more information about these parameters, Setting Up the Advanced Options (on page 223).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

- The *Model Name* is filled automatically. It corresponds to the name of the target variable (`class` for this scenario), followed by the underscore sign ("_") and the name of the data source, minus its file extension (`Census01` for this scenario).

Before generating the model, you can define the **number of clusters** that you want to obtain. These fields allow you to specify how many clusters will be generated by the model. By default the number of clusters is set to 10. The higher the number of segments, the lower the robustness (KR). The lower the number of segments, the lower the information (KI). One should generally start with the default number and then go further with more or less clusters based on the results.

- For **supervised segmentation** (that is to say with a Target), the user chooses the best number of segments, for instance [5;10] which means that 5 to 10 clusters are requested by the user. The engine computes the "best number of clusters" using the metric KI+KR. For instance, you may have 7 clusters.
- For **unsupervised segmentation** (that is to say without Target), the SAP InfiniteInsight® engine chooses the minimum number of clusters, for instance [10;10] which means that 10 clusters are requested by the user.

> **Note**
>
> When you activate the option *Calculate SQL Expressions*, SAP InfiniteInsight® generates an additional cluster that contains the unassigned records. For more details on SQL expressions and unassigned records, see Difference Between Standard Cross Statistics and SQL Expressions (on page 260).

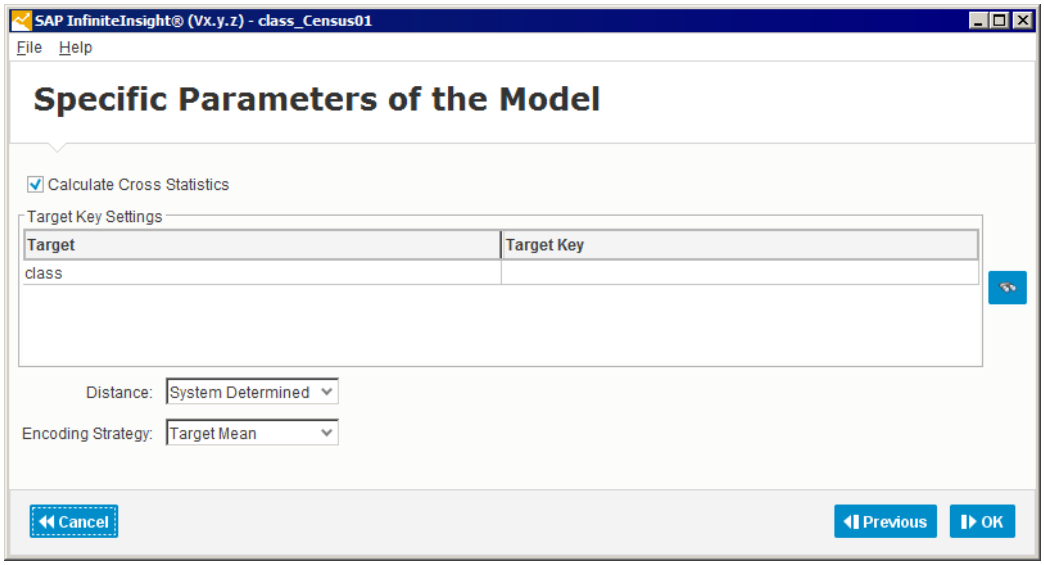
- Choosing to *Calculate SQL Expressions* allows you to see, in the model debriefing, the SQL Expressions used to generate each cluster.

☑ For this Scenario

Keep the default settings.

Setting Up the Advanced Options

The panel *Specific Parameters of the Model* provides you with several options.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

You may calculate the cross-statistics for the model to be generated, define the target key value, choose the distance computing option or choose the encoding strategy.

Calculating the Cross Statistics

This option allows you to visualize the profile of each explanatory variable for each cluster, with respect to their profile for the entire data set.

☑ To Enable the Cross Statistics Calculation

Check the *Calculate Cross Statistics* box.

Defining the Target Key Value

The Set Target Keys value option lists the target variables selected in the Selecting Variables screen and allows you to choose their key value.

☑ To Define the Target Key Value

In the *Target Key* field, enter the key value of the target variable.

Choosing the Distance Computing Method

The **Distance** list allows you to specify the distance used to compare K2C-encoded input data.

☑ To Choose the Distance Computing Method

In the **Distance** drop-down list, select among these options:

- *Chessboard*: maximum of absolute differences between coordinates (LInf).
- *Euclidean*: square root of sum of square differences between coordinates (L2).
- *City Block*: sum of absolute differences between coordinates (L1).
- *System Determined* (default value): Lets the system determine the best distance to be used according to the model build settings.

> **Note**
>
> The current policy is to use LInf either in unsupervised mode or when the clusters SQL expressions have been asked for, and L2 otherwise.

Choosing the Encoding Strategy

The **Encoding Strategy** option refers to the kind of encoding the segmentation engine is expecting from the data encoder of InfiniteInsight®.

☑ To Choose an Encoding Strategy:

Choose among the following options from the drop-down list:

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Option	Description
System Determined	Lets the system select the best encoding according to the model parameters. The *Target Mean* encoding is used for supervised models. Otherwise, variables are encoded using the *Unsupervised* scheme.
Target Mean	*Default value for supervised clustering* Each value of a continuous input variable is replaced by the mean of the target for the segment the value belongs to. Each category of a nominal input variable is replaced by the mean of the target for this category. In case of a nominal target variable, the mean of the target corresponds to the percentage of positive cases of the target variable for the input variable category.
Uniform	Each variable segment is encoded in the range [-1;+1] so that the distribution of the variables is uniform.
Unsupervised	*Default value for unsupervised clustering* A target free strategy. Only segment frequency is used to encode variables.

The following options will only be displayed **when all variables are continuous**:

Option	Description
Natural	This option does not transform the input data.
Min Max	This option encodes the categories of the variable in the range [0,1], where 0 corresponds to the minimum value of the variable and 1 corresponds to the maximum value.
Standard Deviation Normalization	This option performs a normalization based on the variable mean and standard deviation. $\frac{x - Mean}{StdDev}$

Activating the Autosave Option

The panel *Model Autosave* allows you to activate the option that will automatically save the model at the end of the generation process and to set the parameters needed when saving the model.

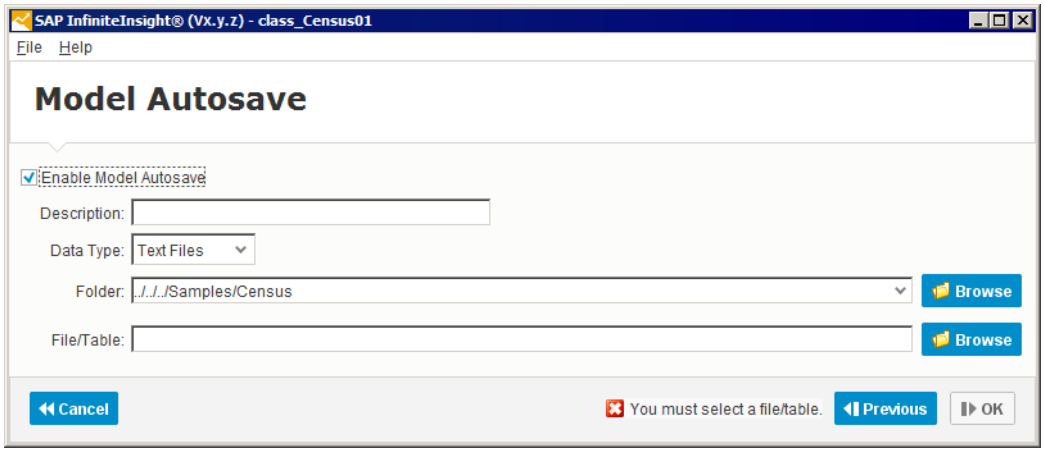
☑ To Activate the Autosave Option

1 In the panel *Summary of Modeling Parameters*, click the *Autosave* button.

The panel *Model Autosave* is displayed.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2 Check the option *Enable Model Autosave*.



3 Set the parameters listed in the following table.

Parameter	Description
Model Name	This field allows you to associate a name with the model. This name will then appear in the list of models to be offered when you open an existing model.
Description	This field allows you to enter the information you want, such as the name of the training data set used, the polynomial degree or the performance indicators obtained. This information could be useful to you later for identifying your model. Note that this description will be used instead of the one entered in the panel *Summary of Modeling Parameters*.
Data Type	this list allows you to select the type of storage in which you want to save your model. The following options are available: ▪ *Text files*, to save the model in a text file. ▪ *Database*, to save the model in a database. ▪ *Flat Memory*, to save the model in the active memory. ▪ *SAS Files*, to save the model in a SAS compatible file for a specified version of SAS and a specified platform (SAS v6 or 7/8 for Windows or UNIX). ▪ *SAS Transport*, to save the model in a generic SAS compatible file.
Folder	Depending upon which option you selected, this field allows you to specify the ODBC source, the memory store or the folder in which you want to save the model.
File/Table	This field allows you to enter the name of the file or table that is to contain the model. The name of the file must contain one of the following format extensions `.txt` (text file in which the data is separated by tabs) or `.csv` (text file in which the data is separated by commas).

4 Click *OK*.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

6.2.2 Step 2 - Generating and Validating the Model

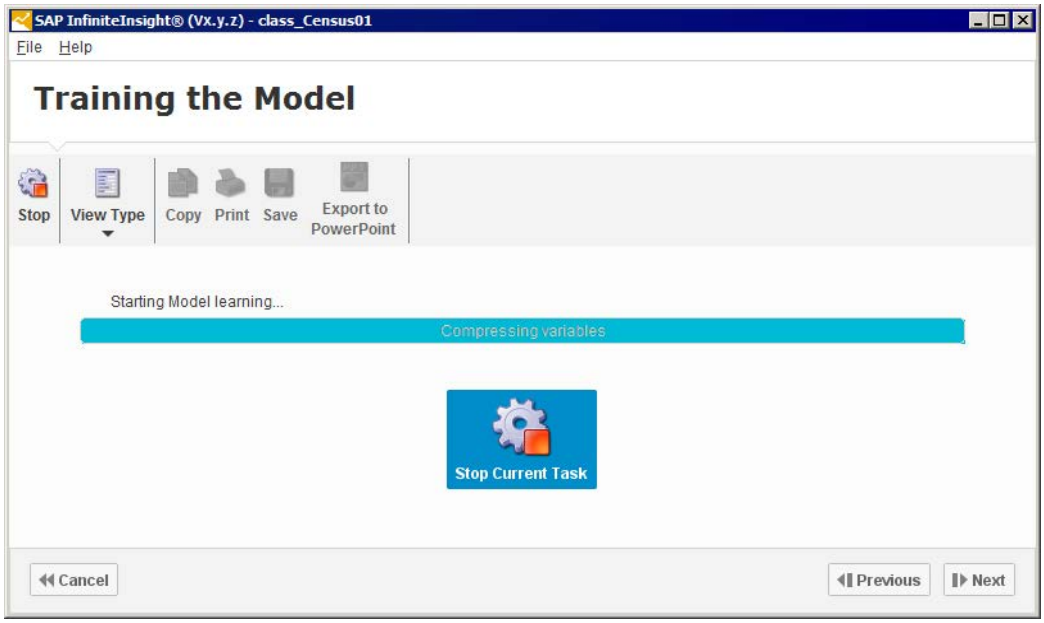
Once the modeling parameters are defined, you can generate the model. Then you must validate its performance using the predictive power (KI) and the prediction confidence (KR):

- If the model is sufficiently powerful, you can analyze the responses that it provides in relation to your business issue (see Step 3 - Analyzing and Understanding the Model Generated (see page 113)), and then apply it to new data sets (see Step 4 - Using the Model (see page 154)).
- Otherwise, you can modify the modeling parameters in such a way that they are better suited to your data set and your business issue, and then generate new, more powerful models.

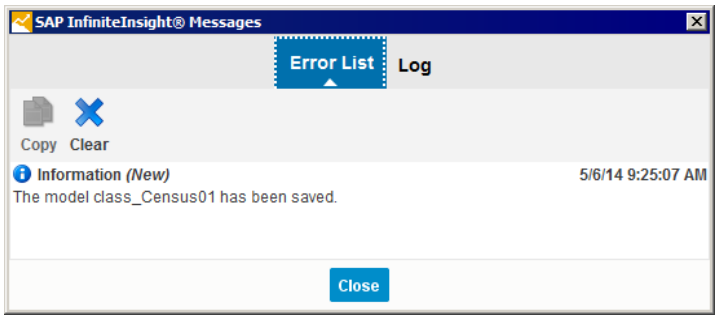
Generating the Model

☑ To Generate the Model

1 On the screen *Specific Parameters of the Model*, click the Generate button.
The screen *Training the Model* appears. The model is being generated. A progress bar allows you to follow the process.



2 If the *Autosave* option has been activated in the panel *Summary of Modeling Parameters*, a message is displayed at the end of the learning process confirming that the model has been saved.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

3 Click *Close*.

4 Once the model has been generated, click *Next* to go to panel **Using the Model**.

Following the Progress of the Generation Process

There is two ways for you to follow the progress of the generation process:

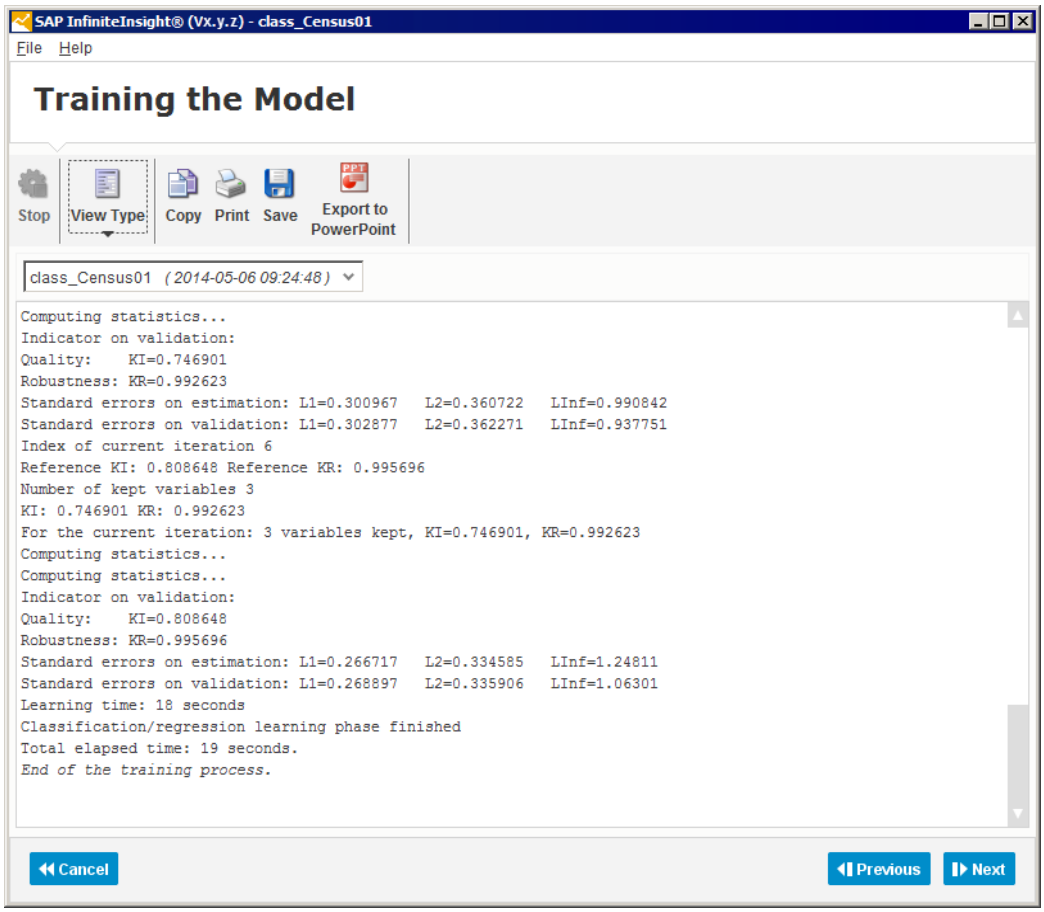
- The **Progress Bar** displays the progression for each step of the process. It is the screen displayed by default.
- The **Detailed Log** displays the details of each step of the process.

☑ To display the Progression Bar

Click *View Type* and select the (*Progression*) button.
The progression bar screen appears.

☑ To Display the Detailed Log

Click *View Type* and select the (*Log*) button.
The following screen appears.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Stop the Learning Process

1 Click the (*Stop Current Task*) button.
2 Click the *Previous* button.
The screen *Summary of Modeling Parameters* appears.
3 Go back to the section **Check Modeling Parameters**.

Validating the Model

Once the model has been generated, you must verify its validity by examining the performance indicators:

- The **predictive power** allows you to evaluate the **explanatory power** of the model, that is, its capacity to explain the target variable when applied to the training data set. A perfect model would possess a predictive power equal to 1 and a completely random model would possess a predictive power equal to 0.
- The **prediction confidence** defines the degree of robustness of the model, that is, its capacity to achieve the same explanatory power when applied to a new data set. In other words, the **degree of robustness** corresponds to the predictive power of the model applied to an application data set.

To see how the predictive power and the prediction confidence are calculated, see Predictive Power, Prediction Confidence and Model Graphs (on page 40).

Beside the *Predictive Power (KI)* and the *Prediction Confidence (KR)*, SAP InfiniteInsight® also provides you with two commonly known indicators:

- the *Classification rate*, in case of a classification model,
- the **Pearson Square Correlation coefficient** (named *R2* in SAP InfiniteInsight®), in case of a regression model.

Both indicators can be used to compare SAP InfiniteInsight® results with results obtained through other data mining tools.

Note

Validation of the model is a critically important phase in the overall process of Data Mining. Always be sure to assign significant importance to the values obtained for the predictive power and the prediction confidence of a model.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

To validate a segmentation model, you can also observe the value of the indicators "frequency" and "target mean" for each of the identified clusters. Specifically, the most interesting clusters of the segmentation model will possess an elevated "frequency" and a "target mean" that deviates from the "target mean" of the entire data set. Note that a segmentation model with a low predictive power may conceal precisely this type of cluster.

To find out how the frequency and target mean for a cluster are calculated, see Understanding the Detailed Description of Clusters.

For this Scenario

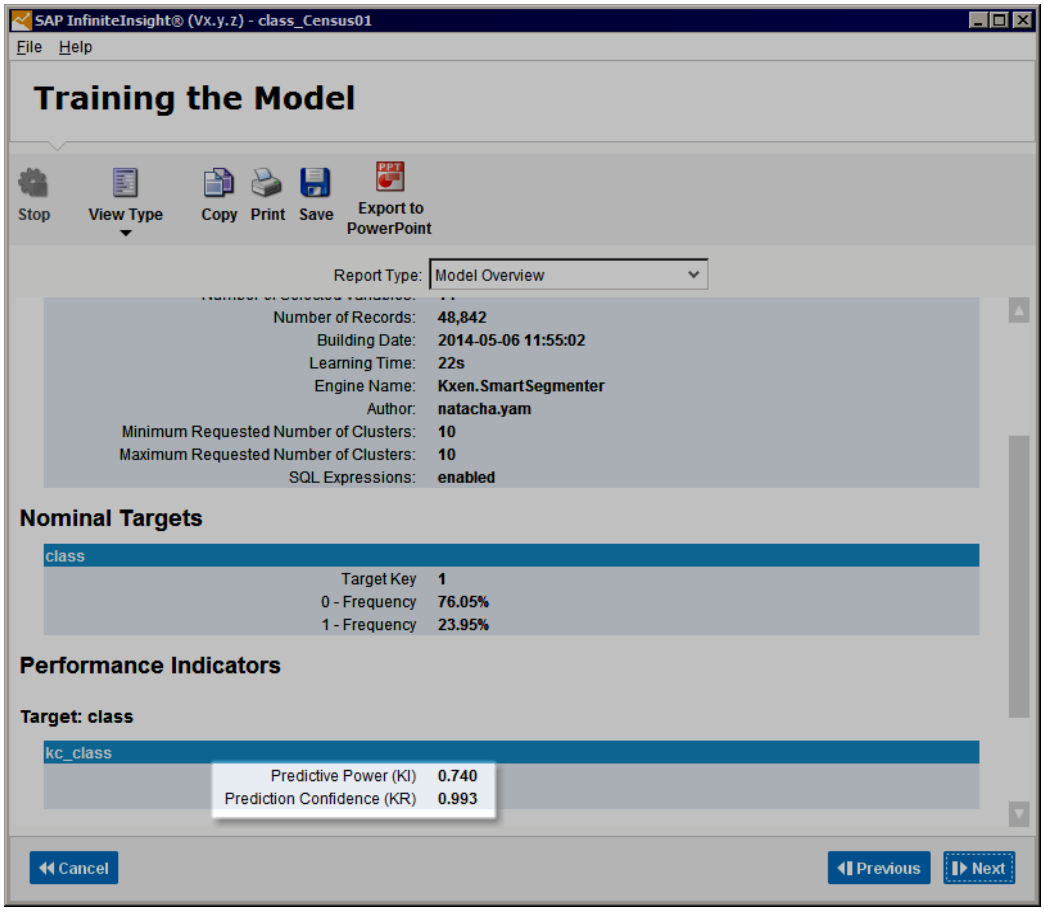
The model generated possesses:

- A predictive power equal to 0.703,
- A prediction confidence equal to 0.987.

The model performs sufficiently well. You do not need to generate another.

To Validate the Model Generated

1 Verify the *Predictive Power (KI)* and the *Prediction Confidence (KR)* of the model. These indicators are highlighted on the following figure.

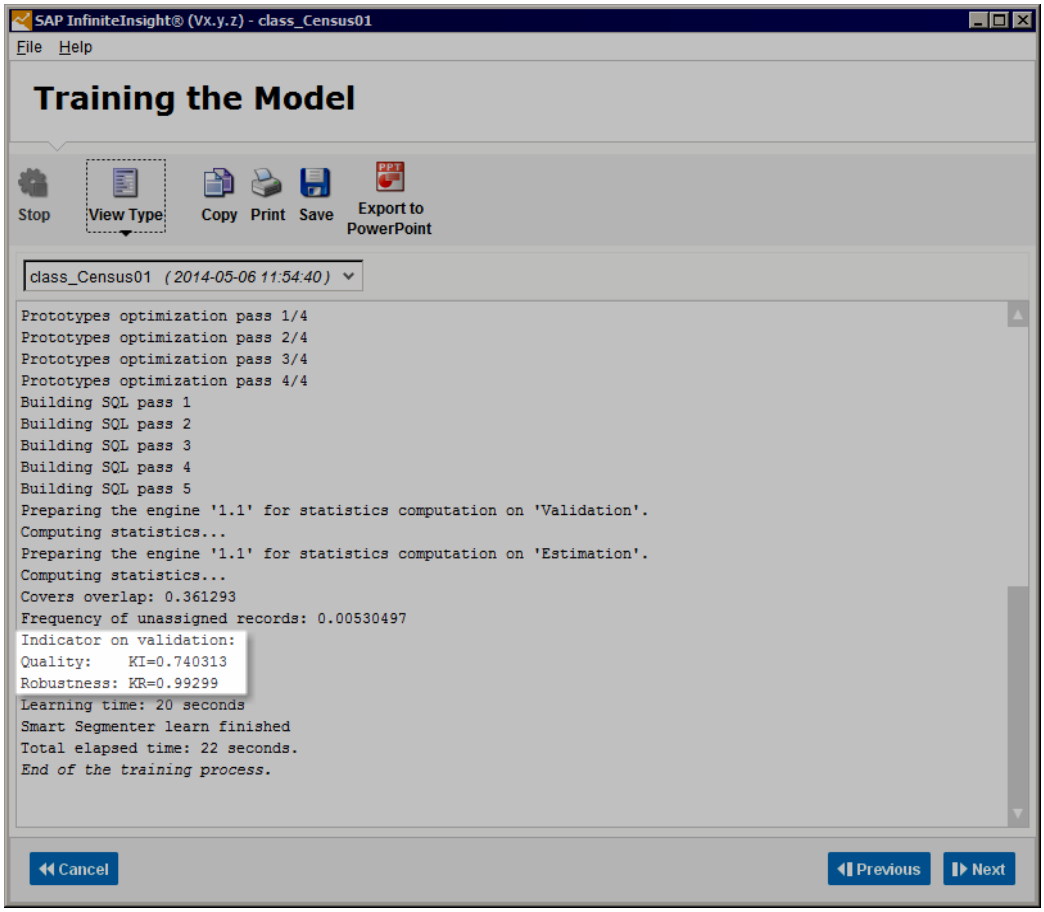


Note

As a general note, other indicators are provided in addition to the predictive power and the prediction confidence during the generation of the model. For example, you could view the *Learning Time* required to generate the model and information on the targets.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2 You can also verify the indicators in the Detailed Log: click the (*Show Detailed Log*) button.
The following screen appears.



3 You can then display the screen *Using the Model*. .
a) If the performance of the model satisfies you, go to **Step 3 - Analyzing and Understanding the Model Generated** (on page 232).
b) Otherwise, go to the procedure To Generate a New Model.

☑ To Generate a New Model

You have two options. On the screen *Training the Model*, you can:

- Either click *Previous* to return to the modeling parameters defined initially.
  Then you can modify the parameters one by one.
- Or click *Cancel* to return to the main screen of Modeling Assistant.
  Then you must redefine all the modeling parameters.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

6.2.3 Step 3 - Analyzing and Understanding the Model Generated

IN THIS CHAPTER

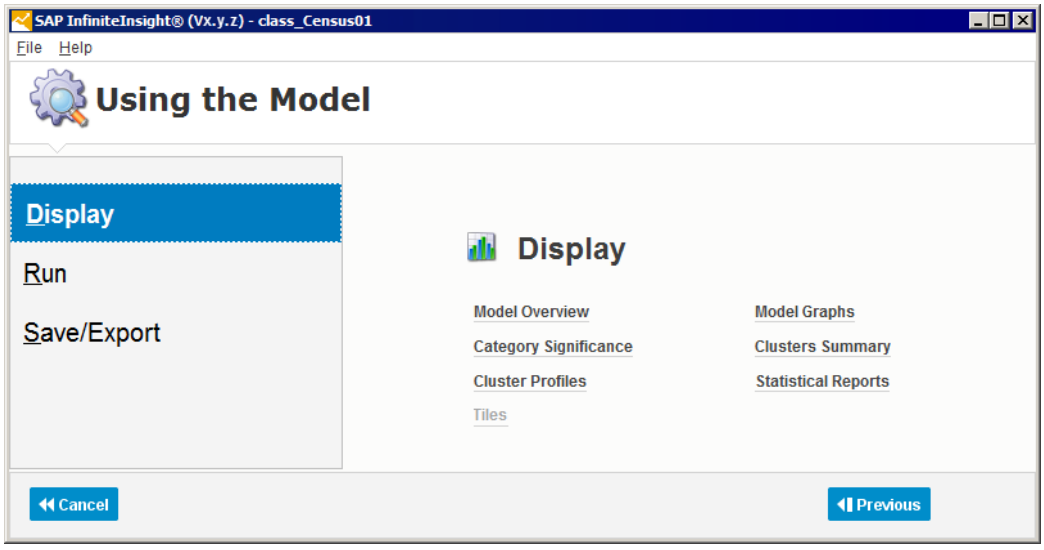
User Menu....232
Model Overview....233
Model Graphs....235
Category Significance....240
Clusters Summary....247
Clusters Profiles....252
Statistical Reports....261

A suite of plotting tools allows you to analyze and understand the model generated:

- The **performance of the model** with respect to a hypothetical perfect model and a random type of model,
- The **characteristics of each of the clusters**,
- The **significance of the various categories of each variable of a cluster** with respect to the target variable (cross statistics).

User Menu

Once the model has been generated, Click *Next*. The screen *Using the Model* appears.



The screen *Using the Model* presents the various options for using the model, that allow you to:

- Display the information relating to the model generated (*Display* section), referring to the model curve plots, plotting of clusters, contributions by variables and the profiles of variables of each cluster.
- Apply the model generated to new data (*Run* section).
- Save the model, or generate the source code (*Save/Export* section).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Model Overview

The screen *Model Overview* displays the same information as the training summary.

Overview

Model	name of the model, created from the target variable name and the data set name
Data Set	name of the data set
Initial Number of Variables	number of variables in the data set
Number of Selected Variables	number of explanatory variables used to build the model
Number of Records:	number of records in the data set
Building Date	date and time when the model was built
Learning Time	total learning time
Engine name	name of the feature used to build the model; *Kxen.KMeans* for a segmentation
Requested Number of Clusters	number of clusters that have been asked for by the user
SQL Expressions	indicates if the SQL expressions for the clusters definitions have been calculated (*Enabled*) or not (*Disabled*)

Nominal Target Variables

For each nominal target:

<TargetVariableName>	name of the target variable for which the statistics are displayed
Target Key	wanted target value
<NonTargetCategory> - Frequency	frequency in percentage of the non-target value in the entire data set
<TargetCategory> - Frequency	frequency in percentage of the wanted target value in the entire data set

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Continuous Targets(Number)

For each continuous target:

<TargetVariableName>	name of the target variable for which the statistics are displayed
Min	minimum value for the target
Max	maximum value for the target
Mean	mean of the target
Standard Deviation	mean of the distance between the target values and the *Mean*

Performance Indicators

For each target variable:

Predictive Power (KI)	For more information on the predictive power, see section **Performance indicators** (on page 39).
Prediction Confidence (KR)	For more information on the prediction confidence, see section **Performance indicators** (on page 39)

Clusters Counts

For each target variable:

Requested Number of Clusters	number of clusters that have been asked for by the user
Effective Number of Clusters	number of clusters found by the model

Model Overview Options

☑ To Copy the Model Overview

Click the (*Copy*) button.
The application copies the HTML code of the screen. You can paste into a word processing or spreadsheet program, a text editor, ...

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Print the Model Overview

1 Click the (*Print*) button situated under the title.
A dialog box appears, allowing you to select the printer to use.
2 Select the printer to use and set other print properties if need be.
3 Click *OK*.
The report is printed.

☑ To Save the Model Overview

Click the (*Save*) button situated under the title.
The file is saved in HTML format.

☑ To Export to PowerPoint

Click the

(*Export to PowerPoint*) button.

Model Graphs

Definition

The **model graphs** allow you to:

- View the **realizable profit** that pertains to your business issue using the model generated.
- **Compare the performance of the model generated** with that of a random type model and that of a hypothetical perfect model.

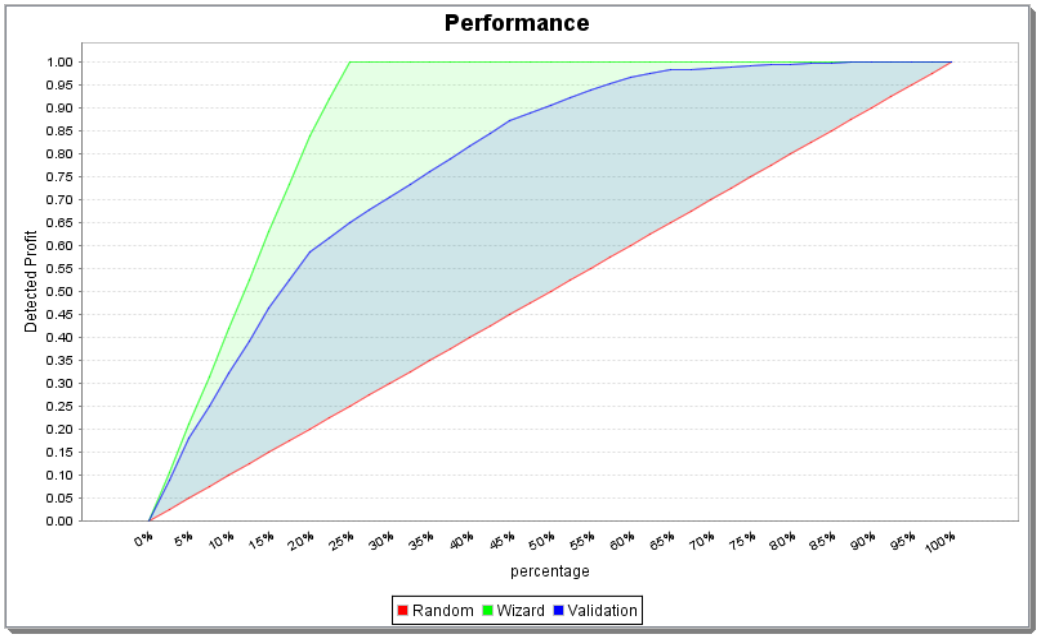
On the plot, for each type of model, the curves represent the realizable profit (Y axis, or ordinate) as a function of the ratio of the observations correctly selected as targets relative to the entire initial data set (X axis, or abscissa).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Displaying the Model Graphs

☑ To Display the Model Graphs

1 On the screen *Using the Model*, click the *Model Graphs* option.
The model graph appears.



The default parameters display the profit curves corresponding to the *Validation* sub-set (blue line), the hypothetical perfect model (*Wizard*; green line) and a random model (*Random*; red line). The default setting for the type of profit parameter is *Detected* profit, and the values of the abscissa are provided in the form of a percentage of the entire data set.

2 When there is more than one target, select the target for which you want to see the curves in the *Models* list.

> **Note**
>
> To each variable corresponds a model. The name of each model is built from the *kc_* prefix and the model target name.

3 Select the viewing options that interest you.
For more information about viewing options, Viewing Options.

Plot Options

☑ To Display the Graphs for the Estimation, Validation and Test Sub-sets

Click *Data Set* and select one of the following options that allow you to switch between :

- the graph for the Validation sub-set ,
- the graphs for all the sub-sets .

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Change the View Type

Click *View Type* and select the desired option.

☑ To Copy the Model Graph

Click the (*Copy*) button and select the desired option.
The application copies the parameters of the plot. You can paste it into a spreadsheet program (such as Excel) and use it to generate a graph.

☑ To Save the Model Graph

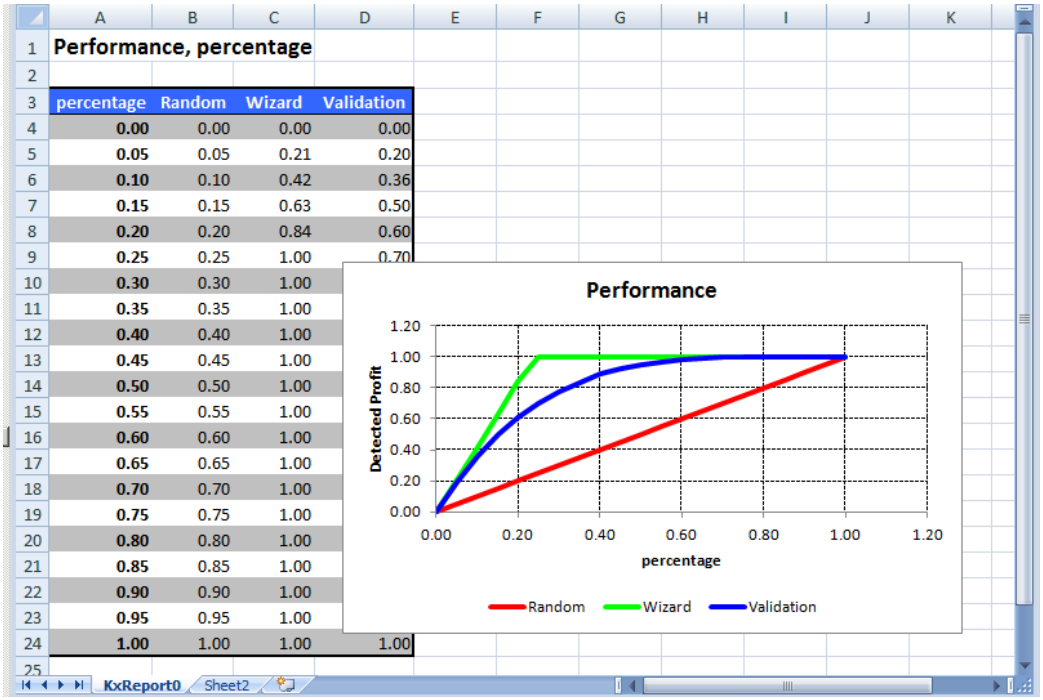
1 Click the (*Save*) button.
A dialog box appears, allowing you to select the file properties.
2 Type a name for your file.
3 Select the destination folder.
4 Click *OK*.
The plot is saved as a PNG formatted image.

☑ To Print the Model Graph

1 Click the (*Print*) button situated under the title.
A dialog box appears, allowing you to select the printer to use.
2 Select the printer to use and set other print properties if need be.
3 Click *OK*.
The report is printed.

☑ To Export the Model Graph to Microsoft Excel

Click the (*Export to Excel*) button situated under the title. An Excel sheet opens containing the model graph you are currently viewing along with its data.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

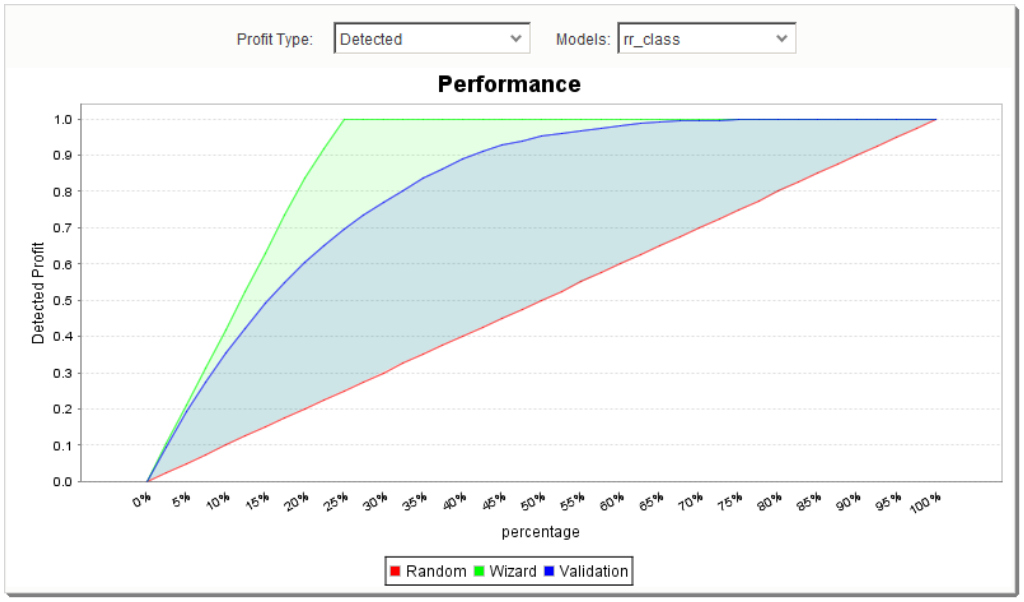
> **Note**
>
> compatible with Excel 2002, 2003, XP and 2007

☑ To Open the Current Graph in a New Window

Click the (*Pin View*) button. The current graph is displayed in a new window.

Understanding the Model Graphs

The following figure represents the model graph produced using the default parameters.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

On the plot, the curves for each type of model represent the profit that may be realized (Y axis), that is, the percentage of observations that belong to the target variable, in relation to the number of observations selected from the entire initial data set (X axis). On the X axis , the observations are sorted in terms of decreasing "score", that is, the decreasing probability that they belong to the target category of the target variable.

In the application scenario, the model curves represent the ratio of prospects likely to respond in a positive manner to your marketing campaign, relative to the entire set of prospects contained in your database.

Detected profit is the default setting for type of profit. Using this type of profit:

- The value "0" is assigned to observations that do not belong to the target category of the target variable,
- The value "1/(frequency of the target variable in the data set)" is assigned to observations that do belong to the target category of the target variable.

The following table describes the three curves represented on the plot created using the default parameters.

The curve...	Represents...	For instance, by selecting...
Wizard (green curve, at the top)	The profit that may be achieved using the **hypothetical perfect model**, that allows one to know with absolute confidence the value of the target variable for each observation of the data set	25% of the observations from your entire data set with the help of a perfect model, 100% of observations belonging to the target category of the target variable are selected. Thus maximum profit is achieved. **Note** These 25% correspond to the proportion of prospects who responded in a positive manner to your marketing campaign, during your test phase. For these prospects, the value of the target variable, or profit, is equal to 1.
Validation (blue curve, in the middle)	The profit that may be achieved using **the model generated by InfiniteInsight® Modeler - Regression/Classification**, that allows one to perform the best possible prediction of the value of the target variable for each observation of the data set	25% of the observations from your initial data set with the help of the model generated, 66.9% of the observations belonging to the target category of the target variable are selected.
Random (red curve, at the bottom)	The profit that may be achieved using a **random model**, that does not allow one to know even a single value of the target variable for each observation of the data set	25% of the initial data set using a random model, 25% belonging to the target category of the target variable are selected.

Predictive Power, Prediction Confidence and Model Graphs

On the model graph plot:

- Of the estimation data set (default plot), the **predictive power** corresponds to "the area found between the curve of the **model generated** and that of the **random model**" divided by "the area found between the curve of the **perfect model** and that of the **random model**". As the curve of the generated model approaches the curve of the perfect model, the value of the predictive power approaches 1.
- Of the estimation, validation and test data sets (select the corresponding option from the list *Data set*, located below the plot), the prediction confidence corresponds to one minus "the area found between the curve of the **estimation data set** and that of the **validation data set**" divided by "the area found between the curve of the **perfect model** and that of the **random model**".

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Category Significance

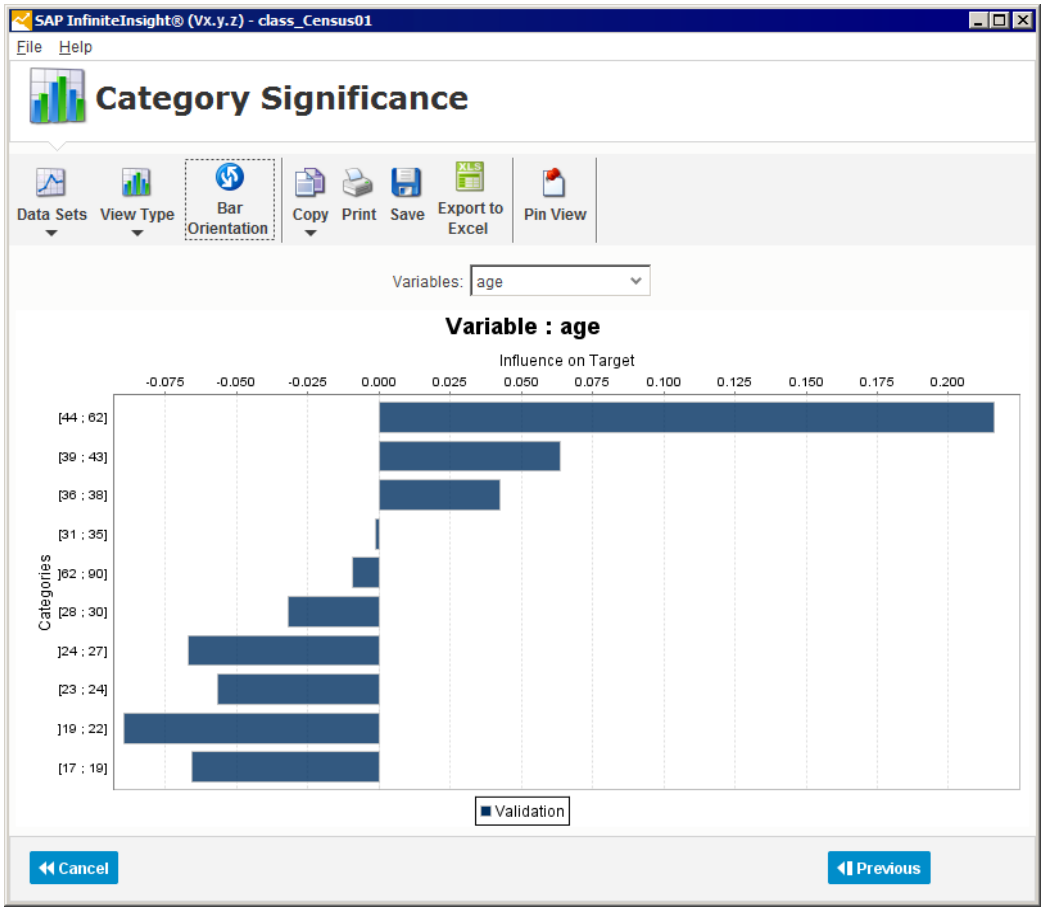
Definition

The *Significance of Categories* plot illustrates the relative significance of the different categories of a given variable with respect to the target variable.

Displaying the Significance of Categories Plot

☑ To Display the Significance of Categories Plot

1 On the screen *Using the Model*, click *Category Significance*.
The plot *Category Significance* appears.



2 In the *Variables* list located above the plot, select the variable for which you want to display the categories.
If your data set contains date or datetime variables, automatically generated variables can appear in the *Variables* list. For more information, refer to section Date and Datetime Variables: Automatically Generated Variables (on page 30).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Note

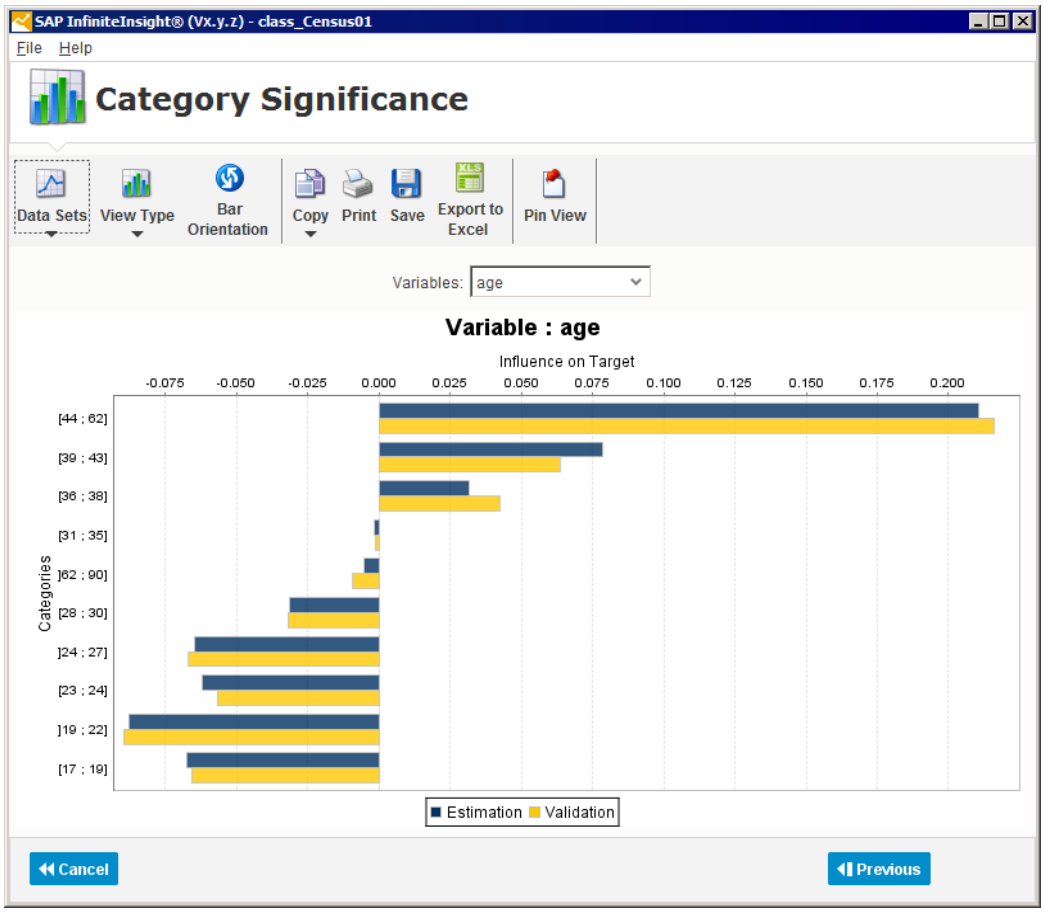
- You can display the relative significance of the categories of a variable directly from the plot *Contributions by Variables*. On the plot *Contributions by Variables*, double-click the bar of the variable which interests you.

- In case no user structure has been defined for a continuous variable, the plot *category significance* displays the categories created automatically using the band count parameter. The number of categories displayed corresponds to the value of the band count parameter. For more information about configuring this parameter, please refer to the section Band Count for Continuous Variables.

Plot Options

☑ To Switch Between "Validation Data Set" and "All Data Sets" Plots

1 Click *Data Sets* and select the (*All Data Sets*) button to display all data sets.
The plot displaying all data sets appears.

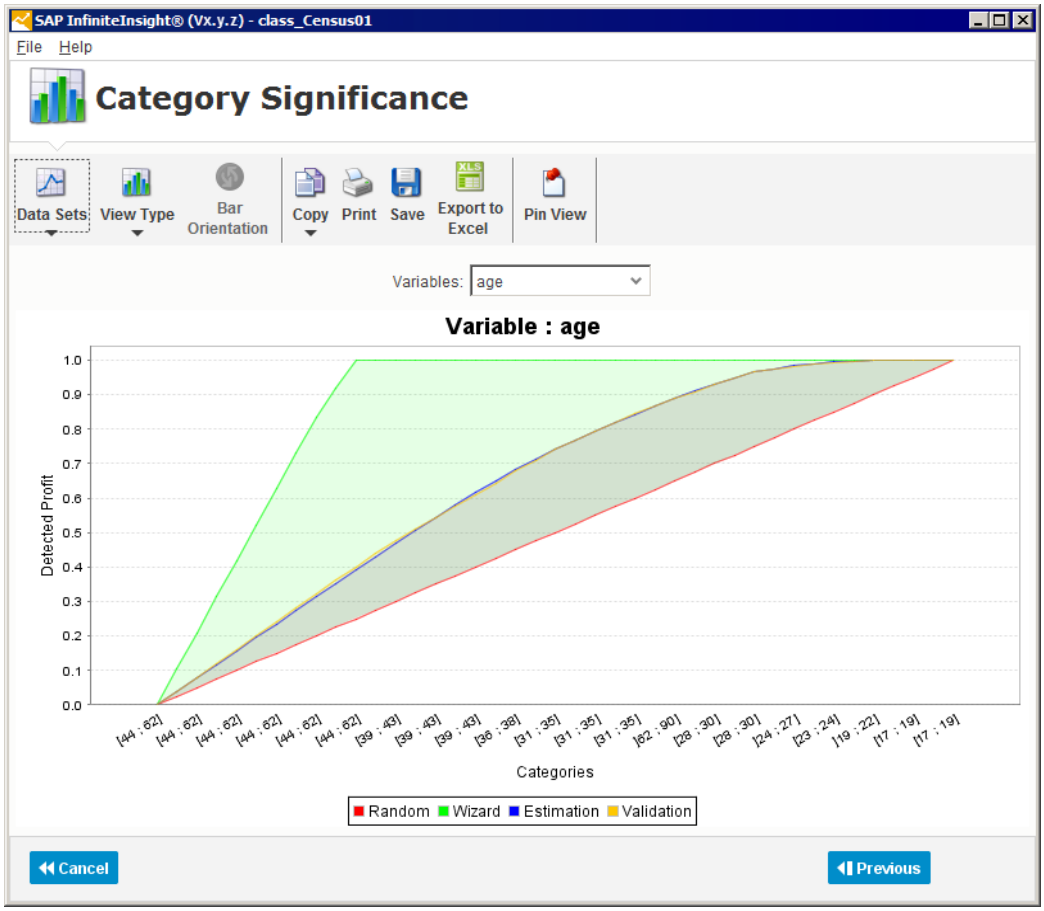


© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2 Click *Data Sets* and select the (*Validation Only*) button to go back to the *Validation Data Set* plot.

☑ To Switch between Curve and Bar Charts

1 Click *View Type* and select the button to display the curve chart.
The curve plot appears.



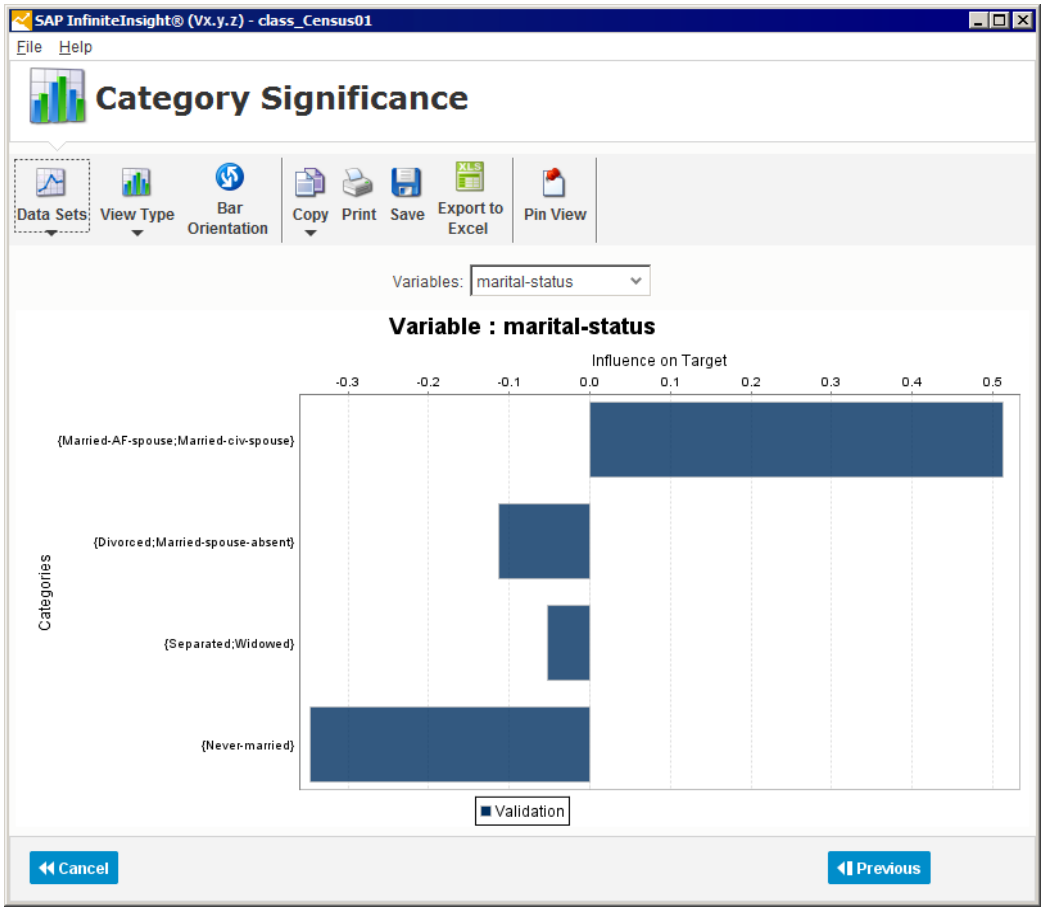
2 Click *View Type* and select the button to go back to the bar chart.

> **Note**
>
> You can combine the different types of plot. For example, you can display *All Datasets* in a curve chart or the *Validation Data Set* in a bar chart.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

For this Scenario

Select the variable *marital-status*, which is the explanatory variable that contributes the most to the target variable *Class*.



This plot presents the effect of the categories of the *marital-status* variable on the target variable.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Variable Categories and Profit

The plot *Category Significance* illustrates the relative significance of the different categories of a given variable with respect to the target variable.

On this type of plot:

- **The higher on the screen one finds a category**, the greater the positive effect on the target category (or hoped-for value) of the target variable. In other words, the higher a category appears on the screen, the more representative that category is of the target category of the target variable.
- **The width and direction of the bar correspond to the profit contributed by that category**. In other words, they correspond to the relationship of that category to the target variable and whether that category has more or less observations belonging to the target category of the target variable. For a given category, a **positive bar** (on the right of 0.0) indicates that the category contains more observations belonging to the target category of the target variable than the mean (calculated on the entire data set). A **negative bar** (on the left of 0.0) indicates that the category contains a lower concentration of target category of the target variable than the mean.

> **Note**
>
> You can display the profit curve for the selected variable by clicking the button (*Display Profit Curve*) located in the tool bar under the title.

The importance of a category depends on both its difference to the target category mean and the number of represented cases. High importance can result from:

- a high discrepancy between the category and the mean of the target category of the target variable
- or a minor discrepancy combined with a large number of records in the category
- or a combination of both.

The width of the bar shows the profit from that category. The positive bars correspond to categories which have more than the mean number from the target category (that is, responders) and the negative bars correspond to categories which have less than the mean number from the target category (that is, responders).

The **Variables** pull down menu allows the selection and graphing of any of the variables in the model.

The tool bar located under the title allows the user to copy the coordinates to the clipboard, print the plot or save it in PNG format.

The values are normalized and their sum always equals to "0". Depending on the chosen profit strategy, or on the continuous target variables value type, you can obtain all positive importances or negative and positive importances.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Axes

The **X-axis** shows the influence of the variable categories on the target. The significance of the different numbers on the X-axis are detailed in the following table.

Number on the X-axis	Indicates that the category has...
positive number	a positive influence on the target
0	no influence on the target (the behavior is the same as the average behavior of the whole population)
negative number	a negative influence on the target

The **Y-axis** displays the **variable categories**. Categories sharing the same effect on the target variable are grouped. They appear as follow `[Category_a;Category_b;Category_c]`. Categories not containing sufficient numbers to provide robust information are grouped in the `KxOther` category. When a variable is associated with too many missing values, the missing values are grouped in the `KxMissing` category. Both categories are created automatically by SAP InfiniteInsight®.

Formulas

`Category Importance = NP*BF/{NC}`

where *NP* is the *Normal Profit*, *BF* is the *Bin Frequency* and *NC* is the *Normalization Constant*.

The calculation of the normalization constant differs by target data type. The calculations for binary and continuous targets are detailed below.

For binary targets, it is calculated as follow: `(Target Frequency) * (1 – Target Frequency])`.

```
It can be approximated for non pathological continuous targets (that is continuous
targets without distribution peak (Dirac)) from:
(Proportion above Median) * (1-Proportion above Median) = 0.5 * (1 - 0.5) = 0.25
```

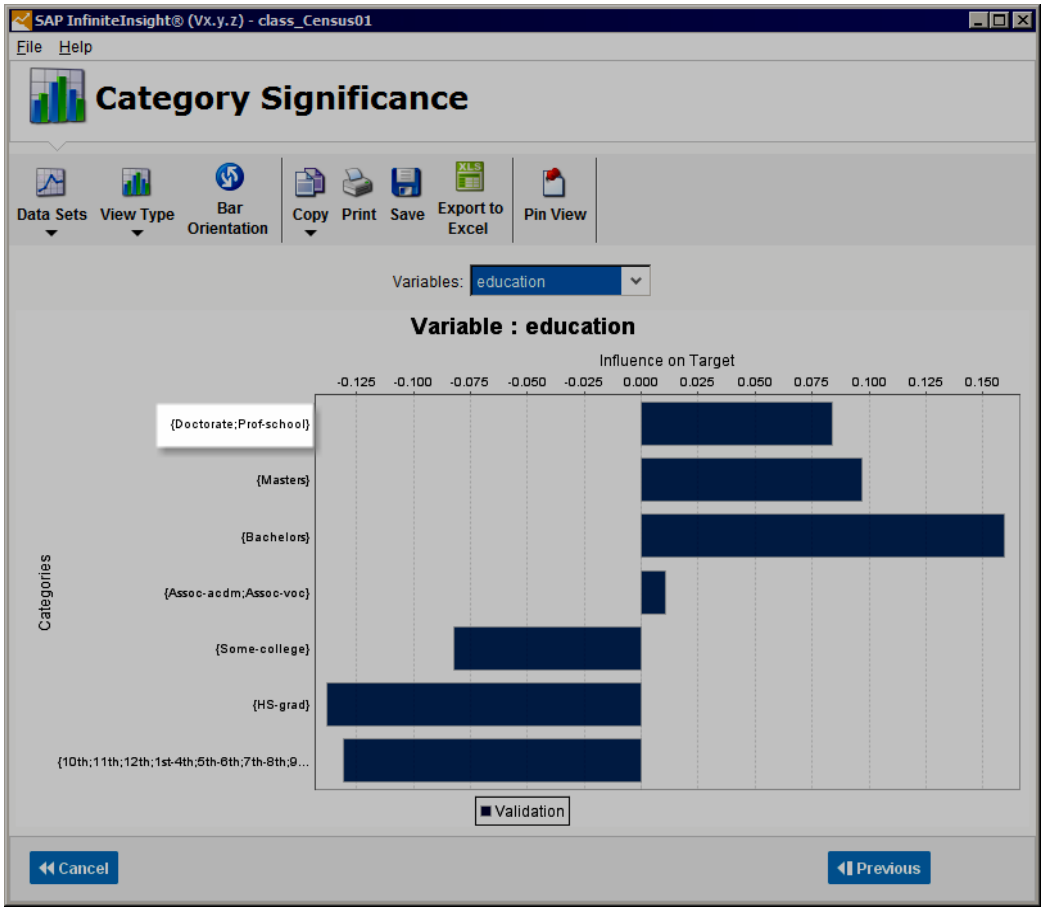

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Grouping Categories

On the plot of details of a variable, categories may appear grouped.

When the option *Optimal Grouping* is enabled, SAP InfiniteInsight® groups those **categories sharing the same effect on the target variable**. For instance, for the variable *education*, the categories *Doctorate* and *Prof-School* are grouped. If the explanatory variable is continuous, SAP InfiniteInsight® identifies the points where behavioral changes occur with respect to the target variable and automatically crops the variable into intervals exhibiting homogeneous behavior with respect to the target.

For more information, please see section Optimal Grouping for all Variables.



When **categories do not contain sufficient numbers to provide robust information**, they are grouped in the *KxOther* category, that is created automatically.

When a variable is associated with **too many missing values**, the missing values are grouped in the *KxMissing* category, that is also created automatically.

To understand the value of the categories *KxOther* and *KxMissing*, consider the following example. The database of corporate customers of a business contains the variable "web address". This variable contains the Web site address of the corporate customers contained in the database. Some companies have a Web site; others do not. In addition, each Web site address is unique. In this case, SAP InfiniteInsight® automatically transforms the "web address" variable into a binary variable with two possible values: *KxOther* (the firm has a Web site) and *KxMissing* (the firm does not have a Web site).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Clusters Summary

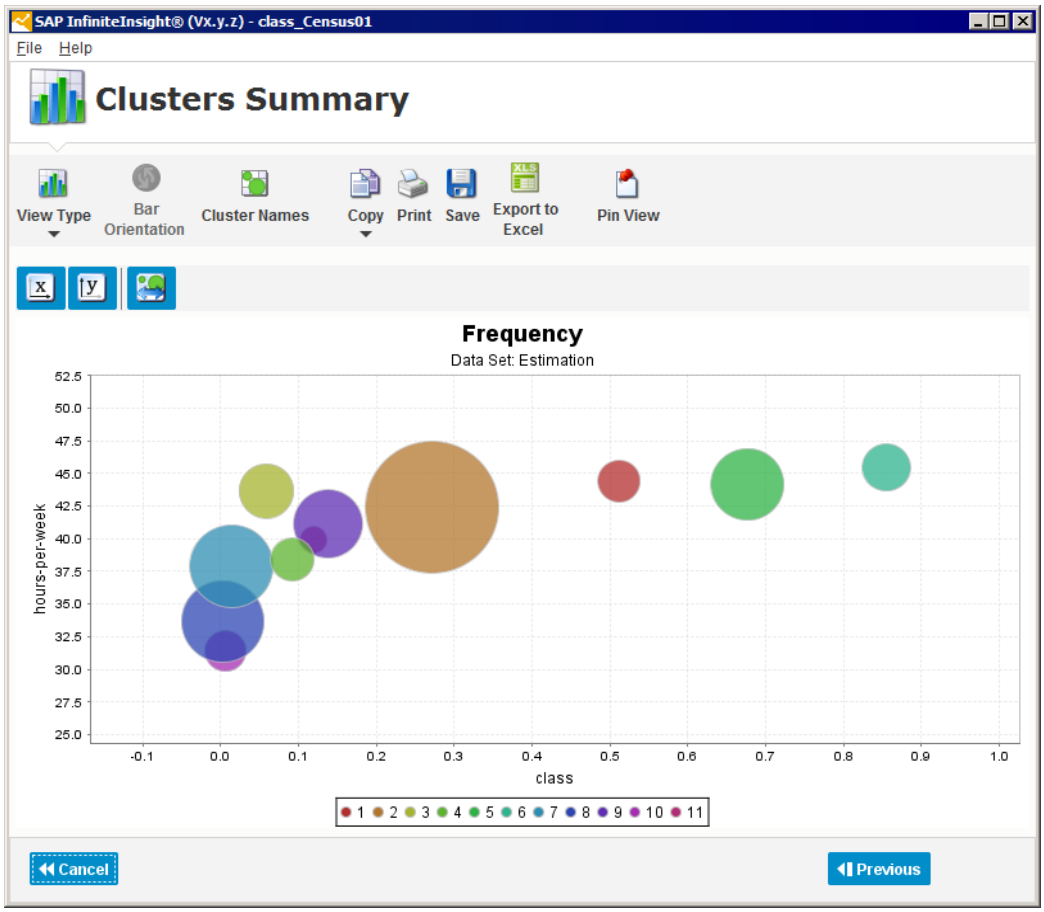
The following types of charts can be displayed:

- **Bubble Charts**
  Bubble charts display the clusters by representing the relationship between three variables.
- **Bar Charts (Cluster Plots)**
  Bar charts displays the three **cluster plots** that allow you to examine:
  - The proportion of observations of the data set contained in each cluster (*Frequencies* plot),
  - The proportion of each cluster relative to the target variable (*Target Means* and *Relative Target Means* plots).

Displaying the Bubble Charts

☑ To Display the Bubble Charts

1 On the screen *Using the Model*, click *Clusters Summary*.
The panel *Clusters Summary* appears.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

2 Use the options to define the variables you want to display in the bubble chart.
The table below lists the available options:

The option...	allows you to...	Note that...
	select the variable to be used in the X-Axis.	only continuous and nominal numerical variables can be used.
	select the variable to be used in the Y-Axis.	only continuous and nominal numerical variables can be used.
	select the variable to be used for the bubble size.	only the variable *Frequency* and the target variable can be used.
	display cluster names.	cluster names can be customized in *Cluster Profiles*.

Understanding the Bubble Charts

The bubble charts allow you to display the clusters representing the relationship between three variables. Thus, a bubble chart can provide 3 pieces of information on each cluster.

In addition, the bubble charts provide a graphical representation of the segmentation, enabling you to easily visualize the clusters. For instance, it can be useful during a presentation.

The X-axis, the Y-axis and the bubble size represent one variable each. You can define the variables to use in the chart. Thus, you can create a bubble chart that separates distinctly the clusters from each other, enabling you to identify the clusters of interest for your marketing campaign.

The figure below represents the relationship between the variables *frequency*, *class* and *capital-gain*.

For instance, results show that the customers listed in *cluster 6* are earning 10,163.4 dollars per year in average (capital-gain: 10,163.4) and represents 5.5% (Frequency: 0.055) of the population listed in the data set. In addition, among these customers, 85.5% (class: 0.885) responded in a positive manner to the test phase of your marketing campaign.

In comparison, *cluster 2* represents the biggest population listed in the data set, namely 25.2 % of the population (Frequency: 0.252), which is around five time bigger than the population listed in *cluster 6*. However, the customers listed in this cluster are earning less the customers listed in *cluster 6*, 147.542 dollars per year in average (capital-gain: 147.542), thus 70% less than *cluster 6*. Moreover, among the customers listed in *cluster 2*, only 27.16% (class: 27.16) responded in a positive manner to the test phase of your marketing campaign.

Consequently, compared to *cluster 2*, *cluster 6* is more interesting because it showed better results to the test phase of your marketing campaign.

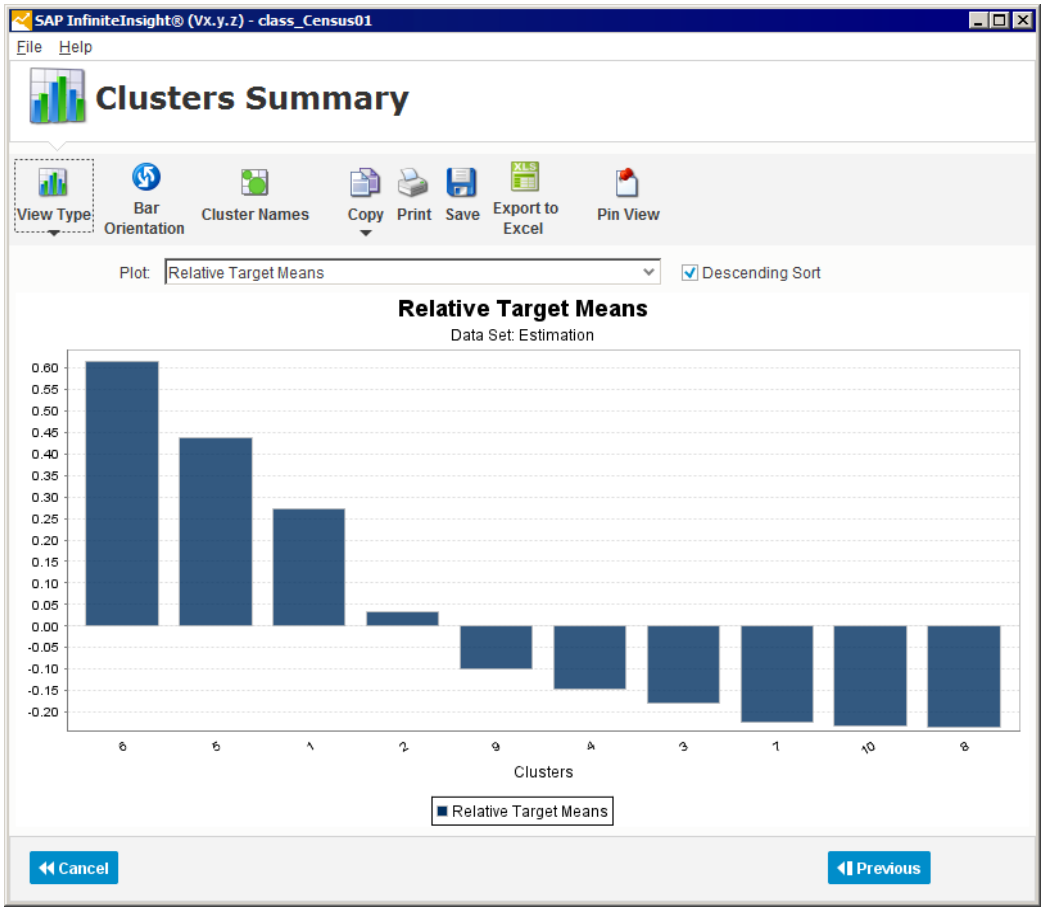
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Displaying Cluster Plots

☑ To Display Cluster Plots

1 On the screen *Using the Model*, click *Clusters Summary*.
The panel *Clusters Summary* appears.

2 Click the button (*View Type*) and select *Bar Chart*.



3 In the *Plot* list, select the type of plot that you want to display.

> **Note**
>
> Select the option *Descending sort* to sort the plot bars in descending order. For instance, on the plot *Relative Target Means*, the descending sort allows quick examination of the most interesting clusters, that is, those which differ most from the mean behavior of the data set taken as a whole.

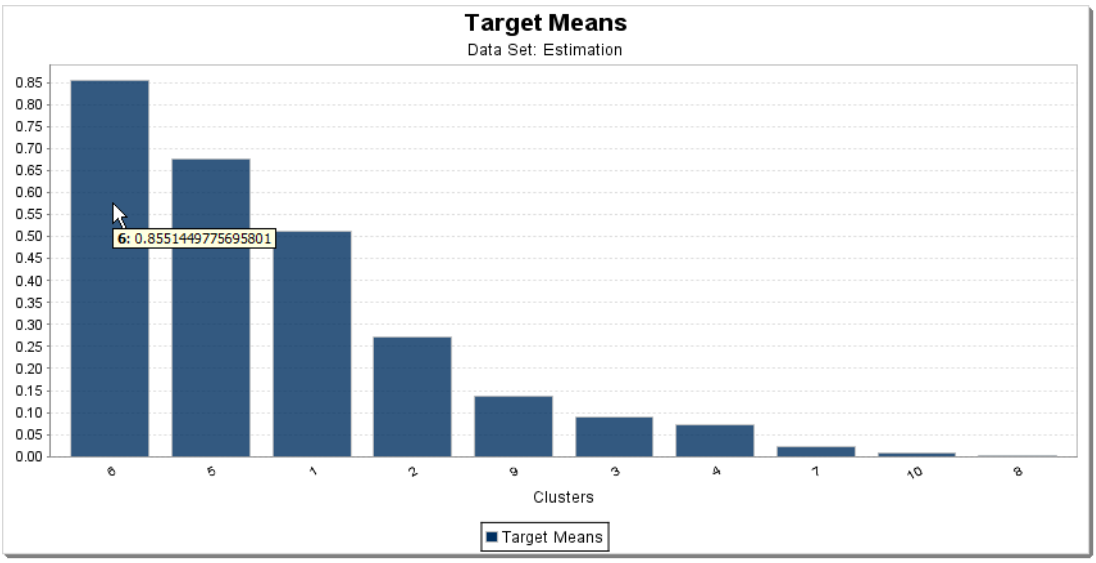
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Understanding Cluster Plots

The Plot "Target Means"

The *Target Means* plot presents the proportion of observations belonging to the target category of the target variable, present in each cluster.

The figure below presents the *Target Means* plot obtained in during this scenario. The bars have been sorted in descending order.



Among the five clusters, *Cluster 6* is the one that has the greatest proportion of observations belonging to the target category of the target variable. In fact, 85.5% of the customers contained in cluster 6 belong to target category *1*, target variable *Class*. In other words, 85.5% of the customers contained in cluster 6 responded in a positive manner to the test phase of your marketing campaign.

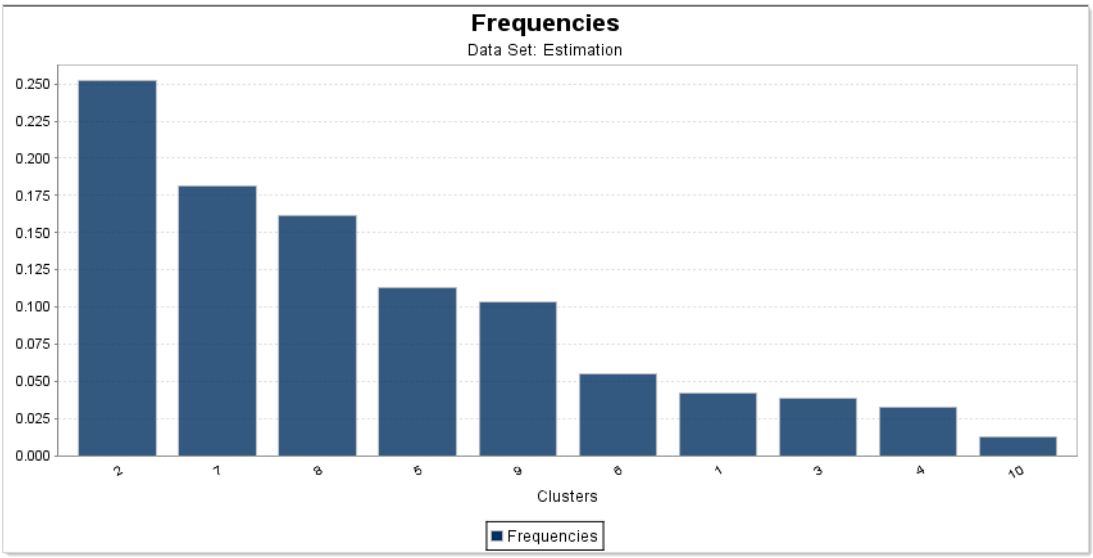
Cluster 8 is the cluster with the lowest density of observations belonging to the target category. Less than 1% of the customers contained in this cluster responded positively to the test phase of your marketing campaign.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

The Plot "Frequencies"

The *Frequencies* plot presents the number of observations contained in each cluster relative to the total number of observations contained in the data set.

The figure below presents the *Frequencies* plot obtained during this scenario. The bars have been sorted in descending order.

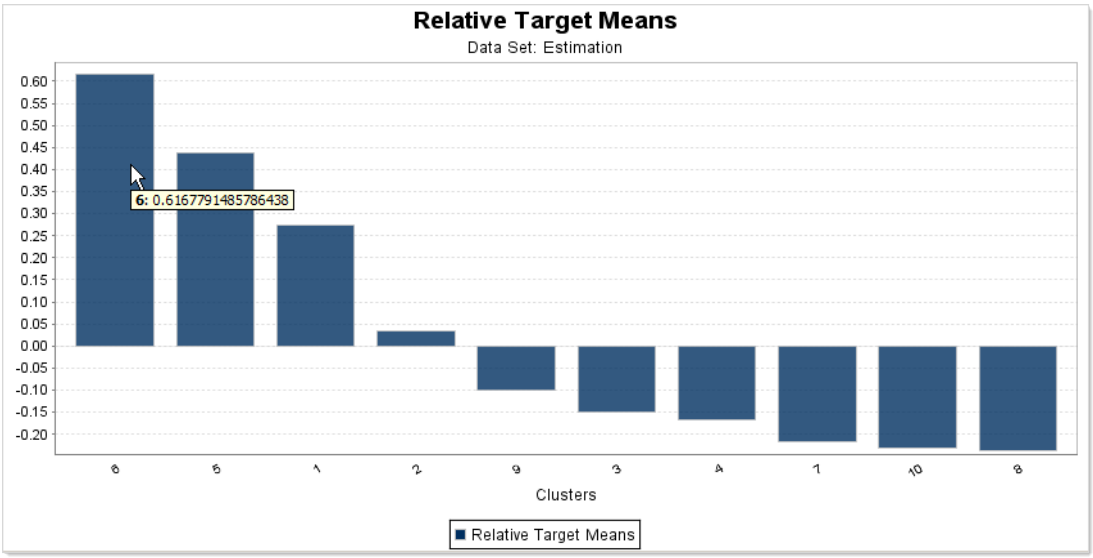


Among the five clusters, *Cluster 2* is the one which contains the greatest number of observations, or 25,2% of the total number of customers contained in the entire data set.

The Plot "Relative Target Means"

Similar to the *Target Means* plot, the *Relative Target Means* plot presents the proportion of observations, for each cluster, belonging to the target category of the target variable. The only difference between the two plots is the scale used on the Y axis. On the *Relative Target Means* plot, the proportion of observations belonging to the target category of the target variable relative to the entire data set is re-expressed. In other words, the 0 value of the Y axis corresponds to the true percentage of observations belonging to the target category of the target variable in relation to the entire data set.

The figure below presents the *Relative Target Means* plot obtained during this scenario. The bars have been sorted in descending order.



 © 2014 SAP AG or an SAP affiliate company. All rights reserved-

Among the ten clusters, *Cluster 6* is the cluster that has the highest proportion of observations belonging to the target category of the target variable. Compared to the entire data set, *Cluster 6* contains 61.6% more customers belonging to the target category *1*, of the target variable *Class*.

Cluster 2 contains less than 3,3% of customers belonging to the target category. In other words, *Cluster 2* has almost the same customer density belonging to the target category as the data set taken as a whole.

Cluster 8 is the cluster with the lowest density of observations belonging to the target category. Compared to the entire data set, *Cluster 8* contains 23.5% fewer of customers belonging to the target category. This cluster therefore has a density of customers belonging to the target category lower than that of the data set.

Clusters Profiles

Cross-Statistics and Variables Profiles

The **clusters profiles** allow you to view for each cluster:

- the profile of each explanatory variable, with respect to their profile over the entire data set,
- the SQL expression of the cluster when they have been calculated.

Variable Profile

The **Variable Profile** indicates the distribution of observations (belonging to a cluster of global data set) within the categories of each variable. In other words, the profile indicates the proportion of observations contained in each of the categories of that variable.

Example of a Variable Profile

The variable "gender" of a data set can be distributed as follows:

- 53% of observations belong to the category "male",
- 47% of observations belong to the category "female".

This distribution corresponds to the **profile** of the variable "gender" over this data set.

Given a cluster A, taken from this data set, the same variable "gender" may be distributed as follows:

- 80% of observations belong to the category "male",
- 20% of observations belong to the category "female".

This distribution corresponds to the **profile** of the variable "gender" over cluster A.

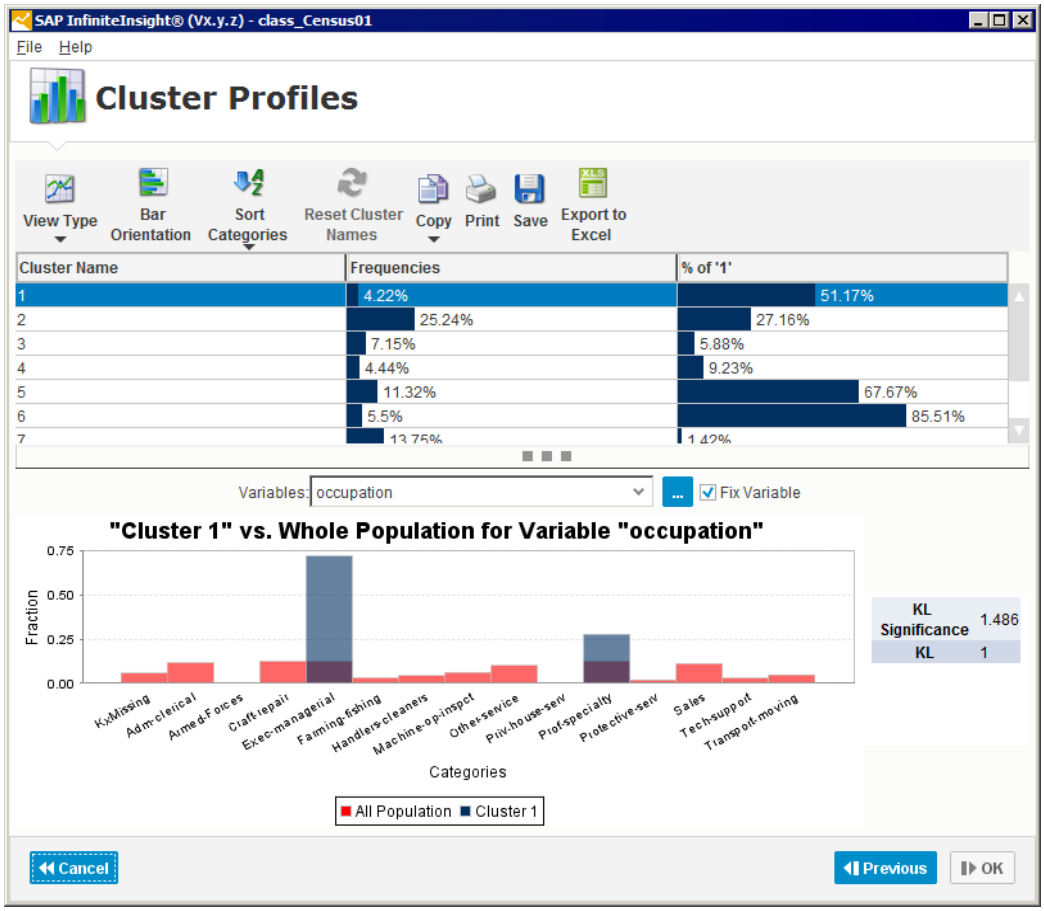
The **clusters profiles** allow you to view and compare the profiles of the variable "gender" over the data set and the clusters taken from this data set.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Displaying Clusters Profiles

☑ To Display Cluster Profiles

1 On the screen *Using the Model*, click *Clusters Profiles*.
The screen *Clusters Profiles* appears.



2 In the table, select the cluster for which you want to view the profile.

> **Note**
>
> If only the variable ranges for a specific cluster are displayed, click the black horizontal bar, a moving cursor is displayed. Drag the cursor down to display the list of clusters.

3 Above the table, from the drop-down list associated with the *Variable* field, select the variable for which you want to see the profile.
The cross statistics will appear in the form of a plot in the lower part of the screen.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Understanding Clusters Profiles

The screen *Cluster Profiles* can be br*OK*en down into three parts:

- In the **upper part**, a drop-down list allows you to select the **variable** for which you want to see the cross statistics. Variables are presented in descending order of the significance of their contribution relative to the target category of the target variable. When a cluster is selected, the variables visible in the drop-down list are sorted according to the difference between their cluster profile and their population profile (the Kullback-Leibler divergence is used to measure this difference). The variable that appears first on the list is the variable exhibiting the greatest difference between its two profiles. This sorted list of variables provides the set of discriminatory variables required to describe a cluster.
- In the **middle part**, a table presents each cluster in a summarized fashion.

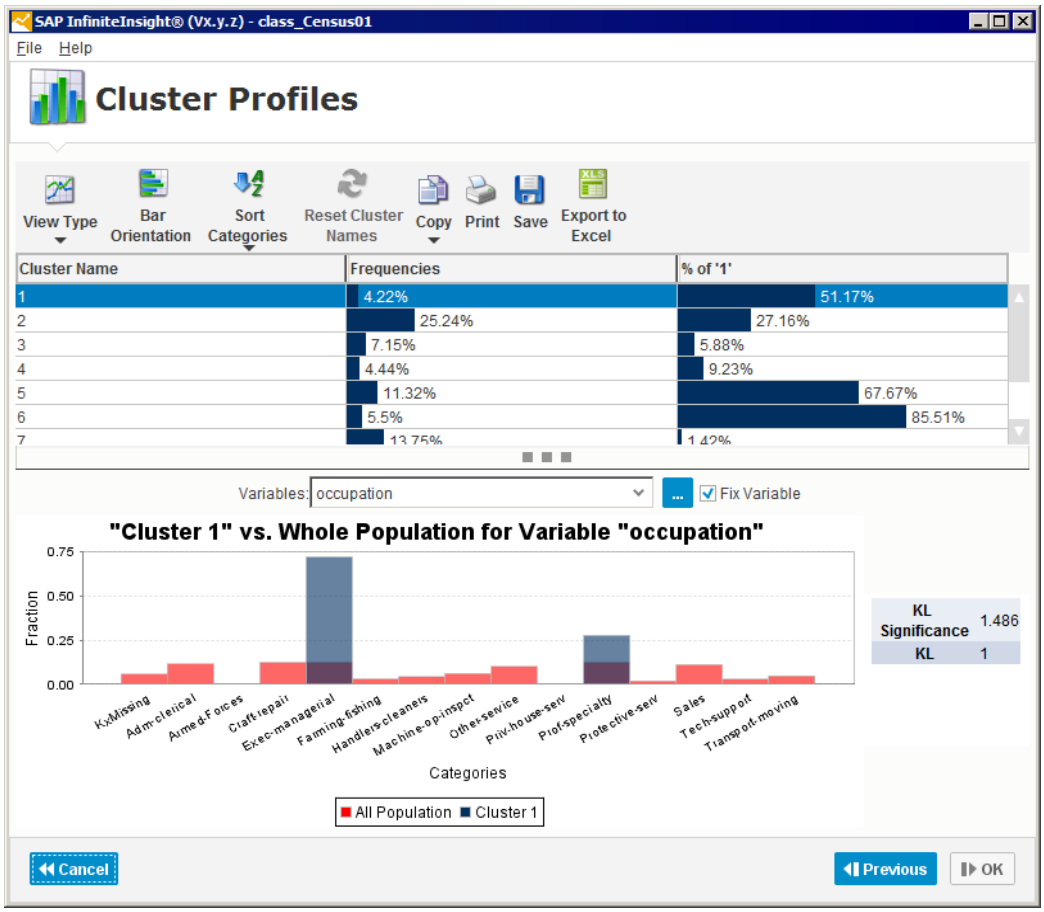
Column ...	Indicates...	For instance...
Cluster Name	The name of the cluster	*Cluster 1*
Frequencies	The number of observations contained in the cluster relative to the total number of observations contained in the data set	The customers contained in cluster 1 represent 4.22% of the total number of customers contained in your entire training data set
% of '1'	The proportion of observations contained in the cluster belonging to the target category of the target variable	51.17% of the customers contained in cluster 1 belong to the target category of the target variable *Class*. In other words, 51.17% of the customers contained in this cluster responded in a positive manner to the test phase of your marketing campaign

This table allows you to select the **cluster** for which you want to view cross statistics.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

- In the **lower part**, a plot presents either the cross statistics corresponding to the cluster and the variables selected, or; when it has been calculated, the SQL expression defining the cluster.

  The figure below presents the screen *Cluster Profiles* which appears as the default plot for this scenario. The plot presents the SQL expression for *cluster 1*.



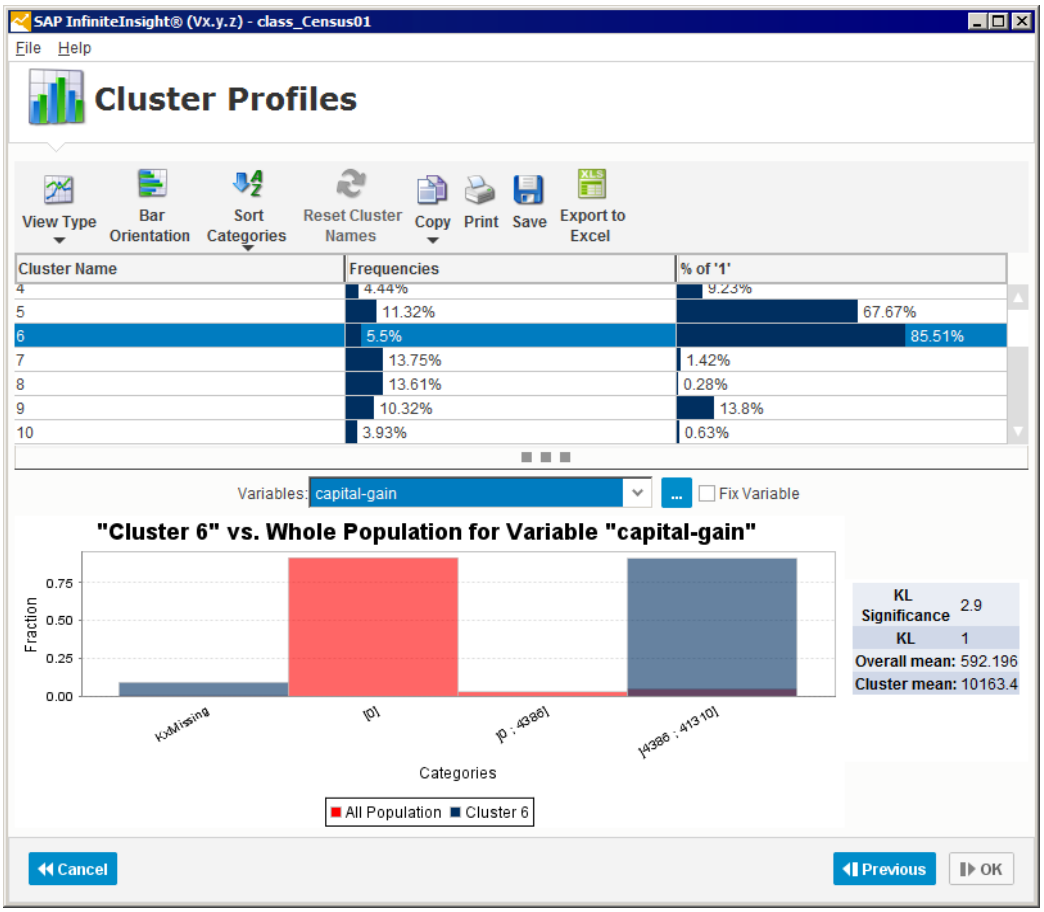
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Cross Statistics Plots

Cross statistics plots contain two curves:

- The **blue** area corresponds to the profile of the variable selected over the cluster selected,
- The **red** area corresponds to the profile of the variable selected over the entire data set.

The figure below presents the *Cross Statistics* obtained in this scenario for *cluster 6*.



© 2014 SAP AG or an SAP affiliate company. All rights reserved-

In the figure above, the table allows you to identify *cluster 6* as the cluster containing the highest density of observations belonging to the target category of the target variable. 88.8% of customers contained in this cluster belong to target category *1* of the target variable *Class*.

The cross statistics plot allows you to view and compare the profiles of the variable capital-gain over the entire data set and over cluster 6. These profiles are repeated in the table below.

Categories of the variable "capital-gain"	Profile over the data set	Profile over cluster 6
KxMissing	1%	6%
[0]	92%	0%
]0 ; 4386]	3%	0%
]4386 ; 41310]	5%	91%

The data distribution of the category *]4650 ; 41310]* makes it clear that the majority of customers contained in cluster 7 realize significant annual capital gains relative to the entire set of customers contained in the data set. In addition, the data distribution over the category *[0]* indicates that the majority of the customers contained in the data set, or 92%, do not realize any annual capital gains, while none of the customers contained in cluster 6 fail to realize some annual capital gain.

Checking the *Fix Variable* box would allow you to compare the profiles of the variable capital-gain for all the segments.

SQL Expressions

The Cross Statistics screen also allows you to visualize the SQL Expression used to define each cluster.

Note

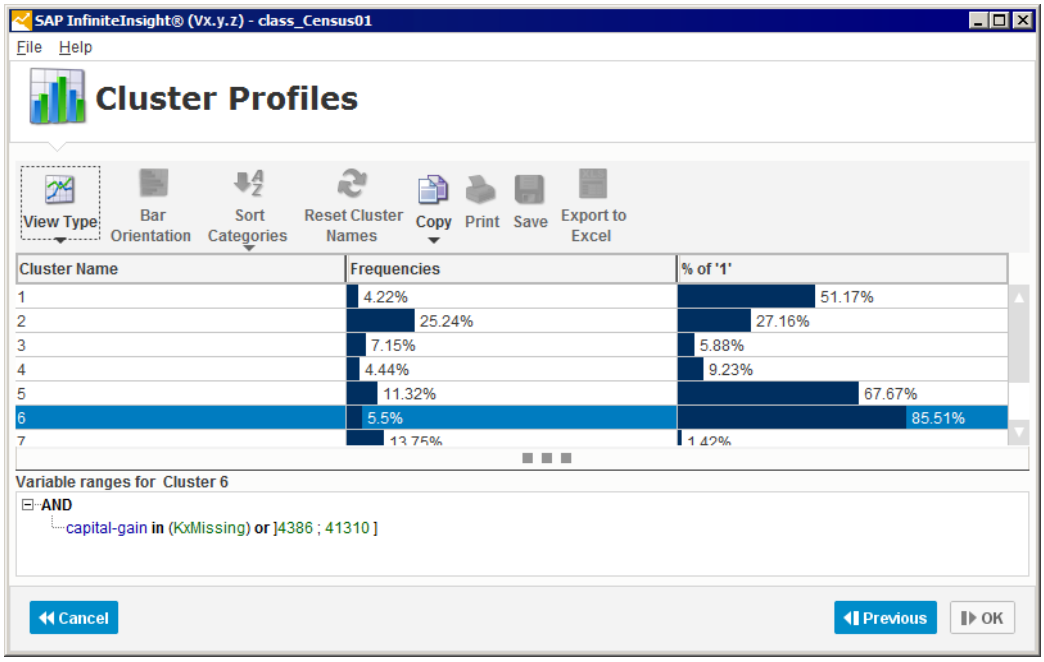
SQL Expressions are only available if you have selected the *Calculate SQL Expressions* option in the Modeling Parameters Advanced Screen before generating your model.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Display SQL Expression for a Cluster

1 Select the cluster in the summary table.
The plot for the selected cluster is displayed.

2 Click *View Type* and select the (*SQL*) button.
The SQL expression replaces the cross statistics plot in the lower part of the screen.



3 Click the small magnifier icon to explore the SQL expression structure.

4 Click *View Type* and select the (*Profiles*) button to go back to the Cross Statistics plot.

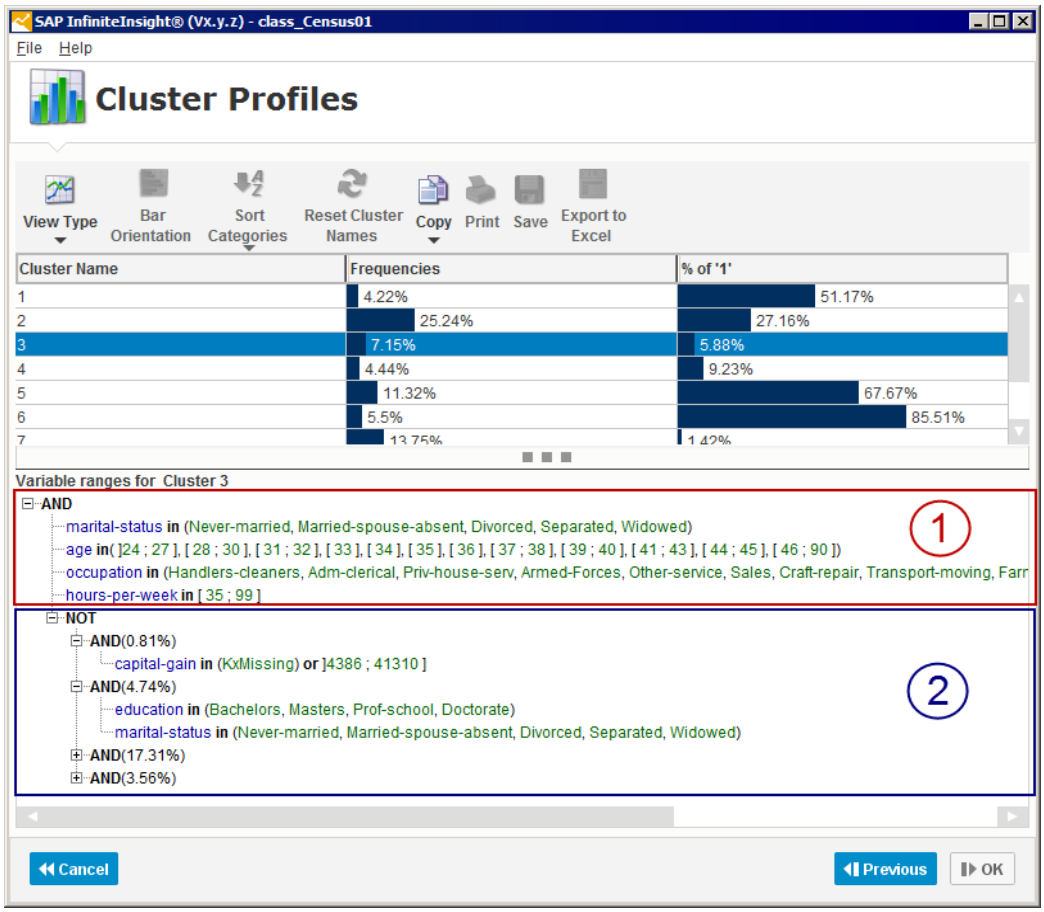
© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Understanding SQL Expressions

The *SQL Expressions* screen can be broken down into two parts:

- In the **upper part**, a table presents each cluster in a summarized fashion. It allows you to select the cluster for which you want to display a SQL expression.
- In the **lower part**, a tree presents the SQL expression corresponding to the selected cluster.

The following schema presents the SQL expression for *Cluster 3*.


SAP InfiniteInsight® (Vx.y.z) - class_Census01
File Help
Cluster Profiles
View Type
Bar Orientation
Sort Categories
Reset Cluster Names
Copy
Print
Save
Export to Excel
Cluster Name
Frequencies
% of '1'
1 4.22% 51.17%
2 25.24% 27.16%
3 7.15% 5.88%
4 4.44% 9.23%
5 11.32% 67.67%
6 5.5% 85.51%
Variable ranges for Cluster 3
AND
marital-status in (Never-married, Married-spouse-absent, Divorced, Separated, Widowed)
age in( ]24 ; 27], [28 ; 30], [31 ; 32], [33], [34], [35], [36], [37 ; 38], [39 ; 40], [41 ; 43], [44 ; 45], [46 ; 90])
occupation in (Handlers-cleaners, Adm-clerical, Priv-house-serv, Armed-Forces, Other-service, Sales, Craft-repair, Transport-moving, Farr
hours-per-week in [35 ; 99]
1
NOT
AND(0.81%)
capital-gain in (KxMissing) or]4386 ; 41310]
AND(4.74%)
education in (Bachelors, Masters, Prof-school, Doctorate)
marital-status in (Never-married, Married-spouse-absent, Divorced, Separated, Widowed)
AND(17.31%)
AND(3.56%)
2
Cancel
Previous
OK

The SQL expression can be br*OK*en down as follow:

- the first part (*1)* defines a cluster of observations where the variables equal the values displayed.
- the second part (*2*) defines clusters of observations that are excluded for the cluster found in part 1. The percentages displayed indicate the proportion of each cluster excluded with respect to the cluster found in part 1.
  In our example, the first excluded cluster corresponds to observations where the capital-gain variable has its value ranges between 4386 excluded and 41310 (*]4386 ; 41310]*). It represents 0.81% of the observations found in part 1.

> **Note**
>
> The clusters are created by applying the SQL expressions in a specific order defined by the engine. If you apply the SQL rules randomly, you may not obtain exactly the same result.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Difference Between Standard Cross Statistics and SQL Expressions

When you ask for SQL expressions, the final segmentation is different from the one without. The goal of SQL is to have easy-to-understand and easy-to-apply segments. SQL expressions are built to describe as much as possible the basic segments (that is the ones you get when you do not ask for SQL). The SQL can be used both to have a better definition/understanding of the clusters and to deploy them on the full database or on new data (which is not usually trivial with other techniques).

The best way to understand the difference between centroid-based clusters and rule-based clusters is to use graphs.

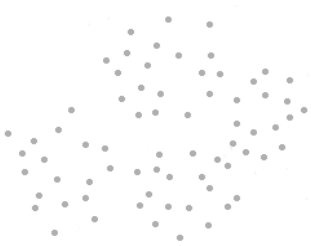
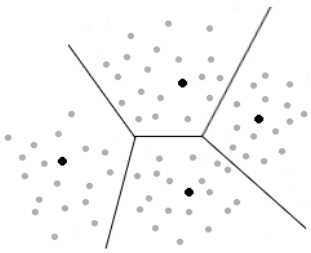
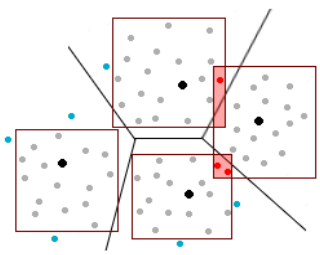
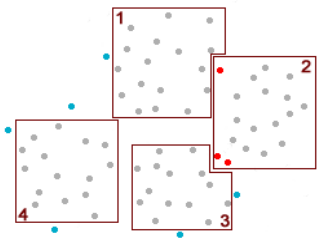
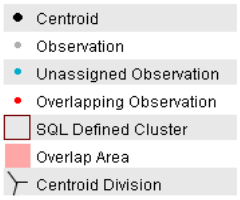
Diagram	Explanation
	This diagram represents a set of observations from a data set.
	To create clusters, InfiniteInsight® Modeler - Segmentation/Clustering engine uses the centroid approach. Centroids are the results of a clustering algorithm, meaning they are the barycenter of the points closest to them. When applying InfiniteInsight® Modeler - Segmentation/Clustering on this data set, the observations are grouped depending on their distance with each centroid. This graph represents the previous data set observations grouped into four clusters. This diagram is known as a Voronoï diagram.
	To create the SQL expressions that define the clusters, InfiniteInsight® Modeler - Segmentation/Clustering engine uses what is called Minimum Description Length (MDL). It means that after creating the initial clusters from the centroid approach then they are reshaped, cut to fit into the smallest possible expression thus trying to find the best compromise between length of the expression and the lost of information. This graph represents the SQL expressions of the clusters (in red) compared with the centroids. You can see on this graph that: ▪ some observations that were in a cluster when using the centroid approach end up in another when using the SQL expressions. ▪ some observations can not be described by the SQL expressions and are left outside the cluster. They are called the *unassigned observations*. ▪ some observations are described by two different SQL expressions, thus appearing in two clusters. This is called the **overlap**.
	This graph presents the final result obtained with SQL expressions. An observation cannot appear in two different clusters, so when there is overlap between clusters, the observation concerned by the overlap is kept in the first cluster created. The second cluster that was also containing the observation is redefined to exclude it. In this schema, the numbers correspond to the order of creation of the clusters. You can see that the observations that were in two clusters are kept in only one. The choice of the cluster in which the overlapping observations are kept depends on the order in which the SQL rules are applied. In this case, the rule defining cluster 2 has been applied before the rules defining the clusters 1 and 3.

Diagram	Explanation
	*Schema Key*

How to decide which segmentation is better ?

As a side effect of the supervision, InfiniteInsight® Modeler - Segmentation/Clustering provides you with a KI and KR. It can be used to compare the two segmentations (especially because the number of segments is the same). If KI does not change significantly, then the one with SQL may be preferred because it is easier to understand. If there is a fall of KI, you may want to stick with the basic segmentation.

KI may not be the thing you want to optimize for segmentation. The target profile of each segment is available in the GUI. Out of the four clusters, maybe one or two are of real interest. In that case you have to focus on these interesting segments and see how they evolve with SQL generation.

Statistical Reports

The *Statistical Reports* provide you with a set of tables that allows you a more detailed debriefing of your model. These reports are grouped in different levels of debriefing:

- the *Descriptive Statistics*, which provides the statistics on the variables, their categories and the data sets, as well as the variables cross-statistics with the target .

  > **Note**
  >
  > If your data set contains date or datetime variables, automatically generated variables will appear in the statistical reports. For more information, refer to section Date and Datetime Variables: Automatically Generated Variables (on page 30).

- the *Model Performance*, in which you will find the model performance indicators, the variables contributions and the score detailed statistics.
- the *Control for Deviations*, which allows you to check the deviations for each variable and each variable category between the validation and test data sets.
- the *Expert Debriefing*, in which you will find more specialized performance indicators, as well as the variables encoding, the excluded variables during model generation and the reason for exclusion and so on.

Statistical Reports Options

A tool bar is provided allowing you to modify how the current report is displayed, to copy the report, to print it, to save it or to export it to Excel.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Display Options

View		This option allows you to display the current report view in the graphical table that can be sorted by column.
		This option allows you to display the current report view as a HTML table.
		Some reports can be displayed as a bar chart. This bar chart can be sorted by ascending or descending values, or by ascending or descending alphabetical order. You can also select which data should be displayed.
		Some reports can be displayed as a pie chart.
		Some reports can be displayed as a line chart.
Sort		When the current report is displayed as a bar chart, this option allows you to change the orientation of the bars (from horizontal to vertical and vice versa).
		This option allows you to display the current report with no sorting.
		This option allows you to sort the current report by ascending values.
		This option allows you to sort the current report by descending values.
		This option allows you to sort the current report by ascending names.
		This option allows you to sort the current report by descending names.
Series	This option allows you to select which columns to display for current report.	

Usage Options

	This option allows you to copy the data from the current view of the displayed report. The data can then be pasted in a text editor, a spreadsheet, a word processing software.	
	This option allows you to print the current view of the selected report depending on the chosen display mode (HTML table, graph, ...).	
		This option allows you to save under different formats (text, html, pdf, rtf) the data from the current view of the selected report.
		This option allows you to save under different formats (text, html, pdf, rtf) the data from all the views of the selected report.
		This option allows you to export to Excel.
		This option allows you to save all reports.
		This option allows you to save the customized style sheet.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

6.2.4 Step 4 - Using the Model

Once generated, a clustering model may be **saved** for later use.

A clustering model may be **applied to additional data sets**. The model thus allows you to assign observations to clusters.

This part presents the option *Applying the model to a new data set* for the InfiniteInsight® Modeler - Segmentation/Clustering feature. The other options for deployment of the clustering models are similar to those proposed for models generated using the InfiniteInsight® Modeler - Regression/Classification feature. For more information about these options, see:

- **Saving the Model**
- **Opening the Model**

Applying the Model to a New Data Set

The currently open model may be applied to additional data sets. The model allows you to determine to which cluster the observations described in these data sets belong.

Constraints of Model Use

In order to apply a model to a data set, the format of the application data set must be identical to that of the training data set used to generate the model. The same target variable, in particular, must be included in both data sets, even if values for the target variable are not contained in the application data set.

Using the Option Direct Apply in the Database

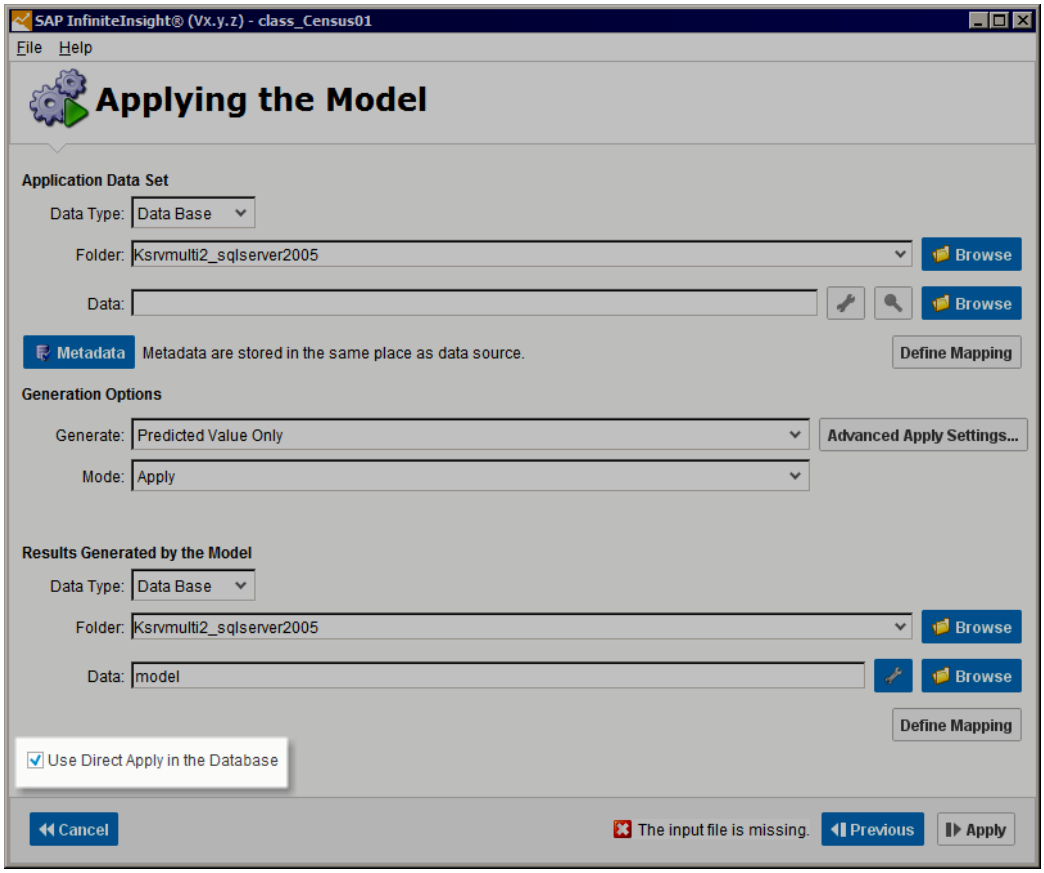
Pre-requisites for the In-database Apply Mode

This optimized scoring mode can be used if all the following conditions are met :

- the apply-in data set (table, view, select statement, data manipulation) and the results data set are tables coming from the same database,
- the model has been computed while at least one physical key variable was defined in SAP InfiniteInsight®,
- there is a valid InfiniteInsight® Scorer license for the database,
- no error has occurred,
- the in-database apply mode is not deactivated,
- granted access to read and write (create table).

☑ To Use the In-database Apply Mode

Check the option **Use the Direct Apply in the Database.**



Advanced Apply Settings

Copy the Weight Variable

This option allows you to add to the output file the weight variable if it had been set during the variable selection of the model.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Copy the Variables

This option allows you to add to the output file one or more variables from the data set.

☑ To Add All the Variables

Check the *All* option.

☑ To Select only Specific Variables

1. Check the *Individual* option.
2. Click the >> button to display the variable selection table.
3. In the *Available* list select the variables you want to add (use the *Ctrl* key to select more than one variable).
4. Click the > button to add the selected variables to the *Selected* list.

User Defined Constant Outputs

This option allows you to add to the output file constants such as the apply date, the data set name, or any other information useful for using the output file.

A user defined constant is made of the following information:

Parameter	Description	Value / Warnings
Visibility	indicates if the constant will appear in the output or not	*checked*: the constant appears in the output *unchecked*: the constant does not appear in the output
Name	the name of the user defined constant	1 The name cannot be the same as the name of an existing variable of the reference data set. 2 If the name is the same as an already existing user defined constant, the new constant will **replace** the previous one
Storage	the constant type (number, string, integer, date, ...)	*number* *string* *integer* *date* *datetime*
Value	the value of the constant	date format: `YYYY-MM-DD` datetime format: `YYYY-MM-DD HH:MM:SS`
Key	indicates if the constant is a key variable or identifier for the record. You can declare multiple keys. They will be built according to the indicated order (1-2-3-...).	*0*: the variable is not an identifier *1*: primary identifier *2*: secondary identifier ...

☑ To Define a Constant

1. Click the *Add* button. A pop-up window opens allowing you to set the constant parameters.
2. In the field *Output Name*, enter the constant name.
3. In the list *Output Storage*, select the constant type.
4. In the field *Output Value,* enter the constant value.
5. Click *OK* to create the constant. The new constant appears in the list. You can choose whether to generate the defined constants or not by checking the *Visibility* box.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Outputs by Cluster Identifier

Top Ranking Centroids Indices

This option allows you to add to the output file the number of the clusters whose centroids are the closest to the current observation. The closest cluster is the one the observation belongs to, its number is displayed in the column *kc_<Target variable>*. The next closest cluster is displayed in the column kc_<Target Variable>_2, and so on until the furthest cluster. you can choose to add all the clusters or only the closest.

☑ To Add All the Clusters

Check the *All* option.

☑ To Add Only the Closest Clusters

1. Check the *Top* option.
2. In the text field enter the number of clusters you want to add (for example, the two, three or four closest).

Top Ranking Centroids Names

This option allows you to add to the output file the names of the clusters whose centroids are the closest to the current observation. The closest cluster is the one the observation belongs to, its name is displayed in the column *kc_name_<Target variable>*. The next closest cluster is displayed in the column *kc_name_<Target Variable>_2*, and so on until the furthest cluster. You can choose to add all the clusters or only the closest.

☑ To Add All the Clusters

Check the *All* option.

☑ To Add Only the Closest Clusters

1. Check the *Top* option.
2. In the text field enter the number of clusters you want to add (for example, the two, three or four closest).

> **Note**
>
> The name of a cluster is by default its number, you can modify this in the column *User Name* of the panel *Clusters Profiles* accessible through the main menu.

Top Ranking Distances

This option allows you to add to the output file the distances of each observation from the clusters centroids. The distance from the closest centroid is displayed int the column *kc_best_dist_<TargetVariable>*, the distance from the second closest centroid is displayed in the column *kc_best_dist_<TargetVariable>_2*, and so on until the furthest centroid. You can add the distances from all centroids or only the shortest.

☑ To Add All the Distances

Check the *All* option.

☑ To Add Only the Shortest Distances

1. Check the *Top* option.
2. In the text field enter the number of distances you want to add (for example, the two, three or four shortest).

> **Note**
>
> When the SQL mode is activated, the notion of nearest cluster does not exist. If a case belongs to a cluster, distance is set to 0. If a case does not belong to a cluster, distance is set to 1.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Probabilities

This option allows you to add to the output file the probabilities that the observation belongs to each cluster. The probability for the observation to belong to the closest cluster is displayed in the column *kc_best_proba_<TargetVariable>*, this probability is usually the highest. The probability for the observation to belong to the second closest cluster is displayed in the column *kc_best_proba_<TargetVariable>_2*, and so on until the furthest cluster. You can add all the probabilities or only the ones corresponding to the closest clusters.

☑ To Add All Probabilities

Check the *All* option.

☑ To Add Only the Probabilities for the Closest Clusters

1. Check the *Top* option.
2. In the text field enter the number of probabilities you want to add (for example, the two, three or four best).

> **Note**
>
> When the SQL mode is activated, the notion of nearest cluster does not exist. If a case belongs to a cluster, probability is set to 1. If a case does not belong to a cluster, probability is set to 0.

Outputs by Cluster Ranks

Distance to Clusters

This option allows you to add to the output file the distance of each observation from the clusters. The distances are generated in the columns named *kc_dist_cluster_<TargetVariable>_<ClusterId>*. For example if the target variable is **Age**, the distance from cluster **1** will appear in the column *kc_dist_cluster_Age_1*.

☑ To Add the Distances from All Clusters

Check the *All* option.

☑ To Select Distances from Specific Clusters

1. Check the *Individual* option.
2. Click the >> button to display the cluster selection table.
3. Check the clusters for which you want to add the distance.

> **Note**
>
> When the SQL mode is activated, the notion of nearest cluster does not exist. If a case belongs to a cluster, distance is set to 0. If a case does not belong to a cluster, distance is set to 1.

Probability for Clusters

This option allows you to add to the output file the probability of each observation to belong to the various clusters. The probabilities are generated in the columns *kc_proba_cluster_<TargetVariable>_<ClusterId>*. For exmaple if the target variable is Age, the probability that the observation belongs to cluster **1** will be displayed in the column *kc_dist_cluster_Age_1*.

☑ To Add the Probabilities for All Clusters

Check the *All* option.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

☑ To Select the Probabilities for Specific Clusters

1. Check the *Individual* option.
2. Click the >> button to display the cluster selection table.
3. Check the clusters for which you want to add the probabilities.

> **Note**
>
> When the SQL mode is activated, the notion of nearest cluster does not exist. If a case belongs to a cluster, probability is set to 1. If a case does not belong to a cluster, probability is set to 0.

Miscellaneous Outputs

Disjunctive Coding

This option allows you to add to the output file the disjunctive coding of the clusters. A column is generated for each cluster and contains either 0 or 1 depending whether the observation belongs to the cluster or not. The columns created are named *kc_disj_<TargetVariable>_<ClusterId>*. For example if the target variable is Age and the model has five clusters, the five following columns will be generated *kc_disj_age_1*, *kc_disj_age_2*, *kc_disj_age_3*, *kc_disj_age_4*, *kc_disj_age_5*.

Target Mean / Target Key Probability

This option allows you to add to the output file:

- for continuous targets:
  - the mean of the target for the cluster containing the observation (displayed in the column *kc_<TargetVariable>_Mean*),
  - the difference with the actual target value if the latter is known for the current observation (displayed in the column *kc_<TargetVariable>_Error*).
- for nominal targets:
  - the proportion of the least frequent category of the target variable (key category) in the cluster containing the current observation (displayed in the column *kc_<TargetVariable>_Mean*).

Types of Results Available

The application of a model to a data set allows you to obtain three types of results:

- The **cluster index** for each observation.
- The **disjunctive encoding** (or **"dummy" coding**) **of the cluster indexes**, which means that, for each cluster, a boolean variable is created indicating whether the current observation belongs to that cluster or not. For a given observation, the value "1" is assigned to the variable corresponding to the cluster containing the observation and the value "0" is assigned to the variables corresponding to the other clusters. The variable names are built according to the following pattern: `kc_<TargetName>_<ClusterIndex>`. Consider, as an example, that you have generated a five-clusters model. When applying this model, SAP InfiniteInsight® creates five variables corresponding to the five generated clusters. For an observation belonging to cluster 3, the result appears as shown below.

KxIndex	class	kc_class	kc_class_1	kc_class_2	kc_class_3	kc_class_4	kc_class_5
15	1	3	0	0	1	0	0

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

- The **target mean for each cluster**, that is, the percentage of observations belonging to the target category of the target variable contained in each cluster.

Depending upon the level of information desired, you can choose to generate:

- Only the **cluster index** to which each observation belongs (*Predicted Value Only* option).
- The **cluster index and the disjunctive encoding of the cluster indexes** (*Cluster Id Disjunctive Coding* option). You can also decide to include in the results file all input variables of the application data set (*Cluster Id Disjunctive Coding (+copy dataset)* option).
- The **cluster index** and the **target mean for each cluster** (*Cluster Id Target Mean* option).

For this Scenario

You will apply the model to the file *Census01.csv*, that you used previously to generate the model.

In the procedure *To Apply the Model to a New Data Set*:

- Select the format *Text files*,
- In the *Generate* field, select the option *Cluster Id Target Mean*,
- Select the folder of your choice in which to save the results file (*Results generated by the model*).

Analyzing the Results of the Application

For this Scenario

Open the results text file in Microsoft Excel, generated when you applied the model to the *Census01.csv* file.

To Open the Model Application Results File

1 Depending upon the format of the results file generated, use *Microsoft Excel* or another application to open the file.
The figure below presents the headings and columns of the results file obtained for this scenario.

	A	B	C	D
1	**KxIndex**	**class**	**kc_clusterId**	**kc_TargetMeanClustId**
2	1	0	3	0,017524
3	2	0	5	0,476401001
4	3	0	3	0,017524
5	4	0	2	0,237075999
6	5	0	5	0,476401001
7	6	0	4	0,308898985
8	7	0	3	0,017524
9	8	1	5	0,476401001
10	9	1	1	0,942696989
11	10	1	1	0,942696989
12	11	1	5	0,476401001
13	12	1	5	0,476401001
14	13	0	3	0,017524
15	14	0	4	0,308898985
16	15	1	2	0,237075999
17	16	0	2	0,237075999

2 You can now analyze the results obtained and use these results of your analysis to make the right decisions.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

Description of the Results File

Depending upon which options you selected, the results file will contain some or all of the following information, in the same order as seen below:

- The **key variable** defined during data description at the model parameter settings step. If your data set did not contain a key variable, the key variable *KxIndex* would have been generated automatically by SAP InfiniteInsight®.
- Possibly, the **target variable** given as known values if the latter appeared in the application data set, as is the case in this scenario.
- The variable *kc_clusterId*, which indicates the number of the cluster to which each observation belongs.
- The variable *kc_TargetMeanClusterId*, which indicates the proportion of observations belonging to the target category of the target variable that are contained in each cluster.
- The variables corresponding to each cluster, and an indication of the **encoding disjunction of the cluster numbers**. The names of these variables correspond to cluster numbers, prefixed by *kc_cluster_*, for example *kc_cluster_1* for cluster 1.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

7 Glossary

A

analytical data management

The **analytical data management** as defined by SAP InfiniteInsight® is made of three elements:

- the **data manipulation functions** offered by SAP InfiniteInsight®, such as filters, join attributes, new attributes (see "attribute" on page 272) computation, aggregates (see page 276), performance indicators (see page 284) definition,
- the **SAP InfiniteInsight® analytical data set methodology**,
- the **meta data management**, which allows storing, sharing and easily re-using the data descriptions.

analytical data set

Tabular representation of data made of lines and columns. Each line represents an “observation”. Roles (see page 287) can be assigned to columns such as “Input”, “skip”, “target” or “weight”.

analytical record

An **analytical record** is a logical view of all **attributes** (see "**attribute**" on page 272) corresponding to an **entity** (see page 277). An analytical record may be decomposed into domains (see page 277) that group attributes related to each other: for example, in CRM, an analytical record can have a demographic domain and a behavioral domain.

antecedent

X is called the antecedent of the rule. The antecedent can be composed of an item (see page 280) or an itemset (see page 280), for example X can be the set {A,B,C}.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

application data set

An application data set (see page 276) is a data set to which you apply a model. This data set contains an unknown target variable (see page 290) for which you want to know the value.

association rule

An association rule is an implication relation of the form X => Y. The rule means: if the attribute X is present in a session, then the attribute Y is present too. Two measures allow qualifying the quality of the rule: the Support (see page 290) and the Confidence (see page 274).

attribute

In computing, an attribute is a specification that defines a property of an object, element, or file.

AUC

The AUC statistic is a rank-based measure of model performance or predictive power calculated as the area under the Receiver Operating Characteristic curve (see "ROC" on page 47). For a simple scoring model with a binary target, this represents the observed probability of a signal (responder) observation having a higher score than a non-signal (non-responder) observation. For individual variables, ordering based on score is replaced by ordering based on the response probability for the variable's categories (for example cluster ID or age range response rates).

authenticated server

Users will be able to communicate to SAP InfiniteInsight® authenticated server only when providing correct password. SAP InfiniteInsight® authenticated server delegates the authentication to Custom built services or Operating System services through PAM (Pluggable Authentication Modules).

autoselection

It is an automated attribute selection.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

B

bin

A bin is a range of values defined by its bounds (upper bound and lower bound). Bins result from a data manipulation activity known as binning. Synonym: range

bipartite graph display / non-bipartite graph display

The bipartite graph display shows two distinct populations of nodes (or node sets) with the links between the two node sets. For example, the first node set could represent clients and the second, products. From this global view, a non-bipartite graph display can be derived to focus on the links between the nodes of a given node set.

bubble chart

A bubble chart is a specific graphical representation in InfiniteInsight® Modeler - Segmentation/Clustering which displays clusters as bubbles. The coordinates of a given bubble are the cluster centroïd values according to two selectable continuous variables. The size of the bubble is plotted according to the frequency of the corresponding cluster.

C

calendar table

A calendar table is used to ease the development of solutions around any business model which involves dates. A common practice is to have a calendar table pre-populated with some or all of the needed information, enabling to accomplish most date related complex tasks with simple database queries.

category

A category is one of the possible values of a discrete variable. A discrete variable is a nominal or ordinal variable. It is the basic element used to code the variable as well as to gather descriptive statistics.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

category significance

The category significance measures the impact a category has on the target.

centroid

Imaginary point inside a polygon whose coordinates are generally those of the polygon center.

chunk (by chunk)

Number of lines of a table that are processed as package.

classification rate

ratio between the number of correctly classified records and the total number of records

confidence

The Confidence of a rule is a measure that indicates the percentage of sessions verifying the consequent among those verifying the antecedent. For instance the number of sessions containing the item D, among the ones containing the itemset {A,B,C}

confusion matrix

The confusion matrix allows visualizing the target values predicted by the model compared with the real values and setting the score above which the observations will be considered as positive, that is the observations for which the target value is the one wanted.

consequent

Y is called the consequent of the rule. The consequent is composed of only one item, for example Y can be the item {D}.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

continuous variable

Continuous variables are variables whose values are numerical, continuous and sortable. Arithmetic operations may be performed on these values, such as determination of their sum or their mean.

contribution

relative importance of each variable in the built model.

correlation

Any measure that quantifies the fact that two variables share the same information. This can be measured by looking at the relative variation of the two variables for different entities. Classical statistics defines linear correlation to compute such a metrics on continuous variables. SAP InfiniteInsight® can compute correlations between variables of different types by looking at the correlation of the codes of both variables in presence of a target.

cross statistics

A method of estimating the accuracy of a classification or regression model. The data set is divided into several parts, with each part in turn used to test a model fitted to the remaining parts.

customized cutting strategy

The customized cutting strategy allows you to define your own data sub-sets. To use this strategy, you must have prepared (before opening SAP InfiniteInsight® features) three sub-sets: the estimation, validation and test sub-sets.

customized profit

Customized profit allows you to define your own profit values, that is, to associate both a cost and a benefit to each value of the target variable.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

cutting strategy

A cutting strategy is a technique that allows decomposition of a training data set into two or three distinct sub-sets:

- An estimation sub-set,
- A validation sub-set,
- A test sub-set.

This cutting allows for cross-validation of the models generated.

D

data aggregation

The process of consolidating data values into a smaller number of values. For example, sales data could be collected on a daily basis and then be totaled to the week level.

data set

a collection of data, usually presented in tabular form, where each column represents a particular variable and where each row is an assignment of values

data source

A data source includes both the source of data itself, such as relational database, a flat-file database, or even a text file, and the connection information necessary for accessing the data.

database

A database is a structured collection of records or data that is stored in a computer system.

descriptive model

A model which allows describing data sets.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

detected profit

Detected profit is the profit type shown as the default. It allows examination of the percentage of observations belonging to the target category of the target variable, that is, the least frequent category, as a function of the proportion of observations selected from the entire data set.

determination coefficient (R2)

ratio between the variability (sum of squares) of the prediction and the variability (sum of squares) of the data.

deviation

Deviation is a measure of difference for interval and ratio variables between the observed value and the mean.

domain

See Analytical Record. The behavioral domain is usually obtained through aggregates per entity on transactional tables.

E

encoding

Encoding is the process of putting a sequence of characters (letters, numbers, punctuation, and certain symbols) into a specialized format for efficient transmission or storage.

engine

The UI and view independent portion of an application, concerned with data manipulation and other fundamental operations independently of how these are eventually represented to the user.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

entity

An **entity** is the object of interest of any analytical task: it can be a customer, a product, a store, and is usually identified with a single identifier that can be used throughout the data repositories. Entities are usually associated with a state model describing the life cycle of such an analytical object of interest.

Note - This is a technical constraint: entities MUST be uniquely identified

error bar

see prediction range

error mean

mean of the difference between predictions and actual values

error standard deviation

dispersion of errors around the actual result

event data set

An event data set should consist of at least:

- an event date such as birthdate or beginning of trial in YYYY/MM/DD format
- a reference id (i.e. customer id), that will be used to join the Events or transactions data with the reference or static customer table previously defined.

excluded variable

actual target

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

explanatory variable

An explanatory variable is a variable that describes your data and which serves to explain a target variable.

Expression Editor

Panel allowing to create fields as complex expressions in the Analytical Data set Editor.

extra-predictable variable

Variable whose values are known for the period that is to be predicted.

F

false positive

incorrect assignments to the signal class

fluctuation

Evolution of the signal that is not stable neither cyclic (InfiniteInsight® Modeler – Time Series).

G

GINI index

The GINI statistic is a measure of predictive power based on the Lorenz curve. It is proportionate to the area between the random line and the Model curve.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

H

horizon-wide MAPE

This quality indicator for the forecasting model is the mean of MAPE values observed over all the training horizon. A value of zero indicates a perfect model while values above 1 indicate bad quality models. A value of 0.09 means that the model takes into account 91% of the signal or , in other words, the forecasting error (model residues) is relatively of 9%.

I

In-database apply

The in-database apply is used to apply a model into a database:

- the proper SQL code is generated for the model,
- the resulting code is then executed as a single SQL request in the database

This avoids extracting the data from the database, and speeds up the writing process of the model outputs.

item

A component of an association rule.

itemset

A group (or a set) of items, is called an itemset.

iteration

An "iteration" is a single loop through a cycle, such as the design-prototype-test cycle.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

K

KL (Kullback-Leibler)

The Kullback-Leibler divergence is used to measure the difference between the cluster profile and the population profile of the variables.

KPI (Key Performance Indicator)

KPIs, or key performance indicators help organizations achieve organizational goals through the definition and measurement of progress. The key indicators are agreed upon by an organization and are indicators which can be measured that will reflect success factors. The KPIs selected must reflect the organization's goals, they must be key to its success, and they must be measurable.

K-S test

K-S is the Kolmogorov-Smirnov statistic applied here as a measure of deviation from uniform response rates across categories of a variable. Kolmogorov-Smirnov is a non-parametric, exact goodness-of-fit statistic based on the maximum deviation between the cumulative and empirical distribution functions.

L

Lift

The Lift of a rule is a measure that indicates the chances of finding the consequent by using the antecedent compared with the chances of randomly finding the consequent. A value greater than 1 indicates that using the antecedent increases your chances to find the consequent.

Lift profit

Lift profit allows examination of the difference between a perfect model and a random model and between the model generated by SAP InfiniteInsight® and a random model.

It represents the ratio between a model and the random model, that is the performance of a model compared to a model that would only allow to select observations at random from your database.

You can thus visualize how much better your model is compared with the random model.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

lower bound

The term lower bound is defined as an element of P which is lesser than or equal to every element of S.

M

maximum error (LInf)

maximum absolute difference between predicted and actual values (upper bound) (Chebyshev distance)

mean

The arithmetic average value of a collection of numeric data.

mean absolute error (L1)

mean of the absolute values of the differences between predictions and actual results, (City block distance or Manhattan distance)

mean absolute percentage error (MAPE)

The MAPE value is the average of the sum of the absolute values of the percentage errors. It measures the accuracy of the model's forecasts and indicates how much the forecasts differ from the real signal value.

mean square error (L2)

square root of the mean of the quadratic errors (Euclidian distance or root mean squared error - RMSE)

metadata

the information about the data itself

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

meta-operator

Operators that are used upon other operators.

missing value

Data values can be missing because they were not measured, not answered, were unknown or were lost.

monotonicity

The direction of variation of a monotonic function does not change.

multiple-instance installation

An installation mode that consists in running several instances on one server in order to divide up the load.

N

nominal variable

Nominal variables are variables whose values are discrete, that is, belong to categories, and are not sortable.

- Nominal variables may be:
- Numerical, meaning that its values are numbers.
- Textual, meaning that its values are character strings.

Important: binary variables are considered nominal variables.

normalize

To transform numerical data and make them fir into a defined interval.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

numeric filter

A digital filter is a system that performs mathematical operations on a sampling, discrete-time signal to reduce or enhance certain aspects of that signal.

O

ordinal variable

Ordinal variables are variables with discrete values, that is, they belong to categories, and they are sortable. Ordinal variables may be:

- numerical, meaning that its values are numbers. They are therefore ordered according to the natural number system (0, 1, 2, and so on).
- textual, meaning that its values are character strings. They are therefore ordered according to alphabetic conventions.

outlier

A data value that does not come from the typical population of data; in other words, extreme values. In a normal distribution, outliers are typically at least 3 standard deviations from the mean.

P

performance indicator (PI)

Performance indicators help organizations achieve organizational goals through the definition and measurement of progress. The purpose of defining PIs is to have a common definition of a metric across multiple projects. A metric like "customer value" could easily be defined in several different ways, leading to confusing or contradictory results from one analysis to the next. Shared PIs ensure consistency across analysts and projects over time. The key indicators are agreed upon by an organization and are indicators which can be measured and will reflect success factors. The PIs selected must reflect the organization's goals, they must be the key to its success, and they must be measurable.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

periodic cutting strategy

The periodic cutting strategy is implemented by following this distribution cycle:

1. Three lines of the initial data set are distributed to the estimation sub-set.
2. One line is distributed to the validation sub-set.
3. One line is distributed to the test sub-set.
4. Distribution begins again at step 1.

pivot

A pivot is a data summarization tool found in data visualization programs. Among other functions, they can automatically sort, count, and total the data stored in one table or spreadsheet and create a second table displaying the summarized data. Pivot tables are also useful for quickly creating cross tabs.

polynomial

A polynomial may be of degree 1, 2, 3 or greater. By defining the polynomial degree, you are defining the degree of complexity of the model.

population

A population is a list of entity identifiers. A population may be defined as list of values. This list can be extracted from a column table: it is then said to be defined in extension, or through a filtering expression from another population: it is then said to be defined in intension.

prediction range

The extreme values for prediction ranges are {TargetMean - (sqrt(TargetVariance)); TargetMean + (sqrt(TargetVariance))}.

predictive model

A model which allows predicting phenomena.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

profit type

A profit type allows calculation of the profit that may be realized using the model. In general, a benefit is associated with the positive (or expected) values of the target variable and a cost is associated with the negative (or unexpected) values.

Q

quality indicator: the predictive power (KI)

The predictive power is the quality indicator of the models generated using SAP InfiniteInsight®. This indicator corresponds to the proportion of information contained in the target variable that the explanatory variables are able to explain.

R

random cutting strategy

The random cutting strategy distributes the data of the initial data set in a random manner between the three sub-sets, estimation, validation and test.

record

The fundamental data structure used for performing data analysis. Also called a table row or example. A typical record would be the structure that contains all relevant information pertinent to one particular customer or account.

robustness

The degree of robustness corresponds to the predictive power of the model applied to an application data set.

robustness indicator: the prediction confidence (KR)

The prediction confidence is the robustness indicator of the models generated using SAP InfiniteInsight®. It indicates the capacity of the model to achieve the same performance when it is applied to a new data set exhibiting the same characteristics as the training data set.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

ROC

The ROC (Receiver Operating Characteristic) graph is derived from signal detection theory. It portrays how well a model discriminates in terms of the tradeoff between sensitivity and specificity, or, in effect, between correct and mistaken detection as the detection threshold is varied.

role

In data modeling, variables (see page 292) may have three roles. They may be:

- Target variables (see page 290),
- Explanatory variables (see page 278),
- Weight variables (see page 293).

root

Terminological morpheme that is used alone as word - root word - or as basic element in a derived word.

Rule mode

A simplified Classification/Regression mode that is used to express the model with rules.

S

score

The numeric evaluation mark in view of a given problem

scorecard

This screen provides you with the coefficients associated to each category for all variables in the model (only in case of a regression model (Classification/Regression)).

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

seasonal

Variations due to calendar events

sensitivity

Sensitivity, which appears on the Y axis, is the proportion of correctly identified signals (true positives) found (out of all true positives on the validation data set).

sequential cutting strategy

The sequential strategy cuts the initial data set into three blocks, corresponding to the usual cutting proportions:

- The lines corresponding to the first 3/5 of the initial data set are distributed as a block to the estimation data set,
- The lines corresponding to the next 1/5 of the initial data set are distributed as a block to the validation data set,
- The lines corresponding to the final 1/5 of the initial data set are distributed as a block to the test data set.

session

A session is identified by a unique key and is composed by one or more transactions.

simulation

Application of a model to only one record

smart variable contributions

The variable contribution in a model while taking into account the variable correlation

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

social network analysis

Social network analysis is used to approach problems such as community identification, diffusion in graphs (product adoption, epidemiology), graph evolution or influence of an individual within a community (leader vs. follower).

standard deviation

The standard deviation is a measure of the dispersion of a collection of numbers.

standardized profit

Standardized profit allows examination of the contribution of the model generated by SAP InfiniteInsight® features relative to a model of random type, that is, in comparison with a model that would only allow to select observations at random from your database.

This profit is used for the plots of variable details, which present the significance of each of the categories of a given variable with respect to the target variable.

statistical report

The Statistical Reports provide you with a set of tables that allows you a more detailed debriefing of your model.

storage

To describe the data, SAP InfiniteInsight® uses five types of storage formats:

- date
- datetime
- number
- integer
- string

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

sub-sampling

Sub-sampling means selecting a part of a whole: if an event cannot be processed as a whole, a limited number of measures have to be taken to represent this event.

support

The support of a rule is a measure that indicates the number of sessions that verify the rule. For instance the number of sessions, that contains the itemset {A,B,C} and the item D.

T

table of data

A table of data is a data set presented in the form of a two-dimensional table.

target key

The target key is the expected value of the target.

target variable

A target variable is the variable that you seek to explain, or for which you want to predict the values in an application data set. It corresponds to your domain-specific business issue.

temporal analytical data set

A **temporal analytical data set** is a special case of analytical data set. It is the product of a **time-stamped population** by an **analytical record**, the result of this operation can be seen as a virtual table containing attributes values associated with identifiers in relation with the time stamp.

In other words: a temporal analytical data set contains 'photos' (or snapshots), of a given list of entities taken at a given time, this time can be different for each entity, and an entity can be associated with several 'photos'.

Note - Analytical data sets are used to train predictive/descriptive models and to apply these models.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

time series

A time series is a sequence of data points, measured typically at successive times, spaced at (often uniform) time intervals.

timeout

A specified period of time that will be allowed to elapse before a specified event is to take place, unless another specified event occurs first.

time-stamped population

A time-stamped population is a list of pairs <identifiers; time stamps>: the semantic meaning of such a construct can be associated with snapshots of the entities and a given time: in general terms, a given entity may be represented at different time stamps in a single time stamped population.

training

Another term for estimating a model's parameters based on the data set at hand.

training data set

A training data set is a data set used for generating a model. By analyzing the training data set, SAP InfiniteInsight® features will generate a model that allows explanation of the target variable, based on the explanatory variables.

transaction

A transaction is defined by:

- a unique key,
- the key of the related session,
- an attribute, called an item.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

true negative

correct assignments to the class of non-signals

true positive

correctly identified signal

U

unassigned record

When creating clusters with SQL expressions, unassigned records are the observations that cannot be described by the SQL expressions and are left outside the cluster.

upper bound

An upper bound of a subset S of some partially ordered set (P, ≤) is an element of P which is greater than or equal to every element of S.

V

variable

A variable corresponds to an attribute which describes the observations stored in your database.

In InfiniteInsight components, a variable is defined by: Type, Storage format, Role.

variable pool

The variable pool is a repository where the user stores the description of the frequently used variables. It is located in the connector store (needs to be associated before, does nothing if not associated). This call stores all the variable information usually edited by the user (standard description: storage; type value; etc, mapping information and structure). The information stored in the pool is retrieved on the next guessed description. The user can also choose to save only the description of a specific variable.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

variable type

There are four types of variables:

- continuous variables
- ordinal variables
- nominal variables
- textual variables

W

weight variable

A weight variable allows one to assign a relative weight to each of the observations it describes, and actively orient the training process. To declare a variable a weight variable results in creating a number of copies of each of the data set observations, proportional to the value they possess for that variable.

© 2014 SAP AG or an SAP affiliate company. All rights reserved-

www.sap.com/contactsap

© 2014 SAP AG or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP AG. The information contained herein may be changed without prior notice.

Some software products marketed by SAP AG and its distributors contain proprietary software components of other software vendors.

National product specifications may vary.

These materials are provided by SAP AG and its affiliated companies ("SAP Group") for informational purposes only, without representation or warranty of any kind, and SAP Group shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP Group products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG in Germany and other countries. Please **see** (**www.sap.com/corporate-en/legal/copyright/index.epx#trademark**) for additional trademark information and notices.